修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　　情報理工学研究科　情報・ネットワーク工学専攻　博士前期課程		
氏　　名	曾根　健太郎	学籍番号	1631077
論 文 題 目	Deep Relational Model の音声信号処理への応用		

要　旨

これまで，音声認識では Hidden Markov Model（HMM）を用いた統計的手法が盛んに研究されている．近年では，HMM と Deep Neural Network（DNN）を用いたハイブリッド方式も提案されており，認識精度を改善することが報告されている．また，音声合成でも HMM を用いた手法が研究されているが，音声認識同様，DNN を用いた手法が合成精度を改善させている．そして，声質変換においても，Gaussian Mixture Model（GMM）に基づく統計的手法よりも，DNN を用いた手法が良い性能を示すことが報告されている．

DNN は多層の隠れ層で構成され，HMM や GMM よりも高い表現能力をもつことから，音声信号処理の様々なタスクにおいて，その性能を発揮している．しかし，DNN は入力から出力への単一方向の関係性しか表現することができない．したがって，DNN を用いた音声認識および合成では，認識器と合成器をそれぞれ独立に構築している．また，声質変換においては，ソース話者からターゲット話者への変換器のみを構築している．

画像の認識および生成において，DNN のような深い構造をもつ生成モデルである Deep Relational Model（DRM）が，画像とそのラベル間の深い双方向の関係性を表現し，画像の認識および生成を行うことが可能であると報告されている．しかし，DRM はバイナリの値のみを扱うため，実数値を扱う必要がある音声信号処理に適していない．

そこで，本論文では，学習コストの削減や精度改善を目的として，DRM を実数値へと拡張した Gaussian-Categorical DRM（GCDRM）および Gaussian-Gaussian DRM（GGDRM）を定義し，音声信号処理へと応用する手法を提案する．そして，音声認識および合成実験結果から GCDRM を用いた提案法が，声質変換実験結果から GGDRM を用いた提案法が，従来法よりも精度を改善することを示す．

平成29年度　修士論文

Deep Relational Modelの音声信号処理への応用

電気通信大学大学院 情報理工学研究科
情報・ネットワーク工学専攻

1631077　曾根 健太郎

指導教員
中鹿 亘 助教
南 泰浩 教授
古賀 久志 准教授

平成30年1月29日

Abstract

In the last decade, statistical approaches using hidden Markov models (HMMs) have been investigated in speech recognition. More recently, hybrid speech recognition systems incorporating deep neural networks (DNNs) with HMMs have achieved the improvements of performances. In speech synthesis, although approaches using HMMs have been explored in speech synthesis, approaches based on DNNs have also improved qualities of the synthesized speech similar to recognition. Also in voice conversion, it is reported that DNN-based systems have outperformed statistical approaches based on gaussian mixture models (GMMs). A DNN, which is a neural network with multiple hidden layers, achieved performance improvements in various tasks of audio signal processing. However, a DNN represents only feedforward dependencies from inputs to outputs. Therefore, DNN-based approaches to speech recognition or synthesis construct only one of either the speech recognition or synthesis system separately. Also, in voice conversion, only the voice conversion system from source speakers to target speakers is constructed. When it comes to the domain of binary-valued image classification and generation, it is reported that a deep relational model (DRM), which contains multiple hidden layers similar to DNNs, has abilities to classify and generate images by representing deep bidirectional relationships between images and class labels. However, since the DRM handles only binary values, it is not suitable for audio signal processing that needs to handle real values. Thus, in this paper, in order to reduce training costs and improve performances, we define a Gaussian-Categorical DRM (GCDRM) and Gaussian-Gaussian DRM (GGDRM) to apply the conventional DRM for the domain of real-valued data, and propose GCDRM-based and GGDRM-based approaches to audio signal processing. Experimental results in speech recognition and synthesis show that the GCDRM-based systems outperform the DNN-based systems. The results in voice conversion also confirm that the GGDRM-based systems perform better the best.

概要

これまで，音声認識では Hidden Markov Model（HMM）を用いた統計的手法が盛んに研究されている．近年では，HMM と Deep Neural Network（DNN）を用いたハイブリッド方式も提案されており，認識精度を改善することが報告されている．また，音声合成でも HMM を用いた手法が研究されているが，音声認識同様，DNN を用いた手法が合成精度を改善させている．そして，声質変換においても，Gaussian Mixture Model（GMM）に基づく統計的手法よりも，DNN を用いた手法が良い性能を示すことが報告されている．DNN は多層の隠れ層で構成され，HMM や GMM よりも高い表現能力をもつことから，音声信号処理の様々なタスクにおいて，その性能を発揮している．しかし，DNN は入力から出力への単一方向の関係性しか表現することができない．したがって，DNN を用いた音声認識および合成では，認識器と合成器をそれぞれ独立に構築している．また，声質変換においては，ソース話者からターゲット話者への変換器のみを構築している．画像の認識および生成において，DNN のような深い構造をもつ生成モデルである Deep Relational Model（DRM）が，画像とそのラベル間の深い双方向の関係性を表現し，画像の認識および生成を行うことが可能であると報告されている．しかし，DRM はバイナリの値のみを扱うため，実数値を扱う必要がある音声信号処理に適していない．そこで，本論文では，学習コストの削減や精度改善を目的として，DRM を実数値へと拡張した Gaussian-Categorical DRM（GCDRM）および Gaussian-Gaussian DRM（GGDRM）を定義し，音声信号処理へと応用する手法を提案する．そして，音声認識および合成実験結果から GCDRM を用いた提案法が，声質変換実験結果から GGDRM を用いた提案法が，従来法よりも精度を改善することを示す．

目 次

第1章

序論

1.1 研究背景と研究目的

音声は言語情報を伝達する声である．人間にとって，情報の多くは自然言語によって表現され，自然言語は音声によって手軽にやりとりすることができる．音声は人間にとって非常に能率の良い情報伝達の方法であるから，人間・コンピュータ間の情報伝達も音声によって行うことができれば便利である．音声言語による人間・コンピュータ間のやりとりを可能にするヒューマンマシンインターフェースをはじめとする様々な音声言語システムを実現するためには，まず音声から言語情報を正確に抽出する必要があり，音声認識を行わなければならない．そして，音声によってコンピュータから人間への情報伝達を行う場合は，言語情報から人工的に音声を生成する必要があり，音声合成を行わなければならない．

近年では我々の日常生活においても，音声信号処理を応用した様々なシステムが広く用いられている．例えば，音声認識を応用したシステムとして，Apple が提供するアプリケーションである Siri があげられる．ユーザは Siri を搭載した端末に話かけるだけで，端末の操作やメールの本文作成を行うことができる．また，Siri がユーザに応答する際には，音声合成によって人工的に生成された音声が出力される．

これまで，音声認識では Hidden Markov Model（HMM）を用いた統計的手法 [1] が研究されている．HMM は入力となる音声信号の特徴量と，自身の内部状態との関係を Gaussian Mixture Model（GMM）を用いて結合分布として表現する（GMM-HMM）．しかし，結合分布をモデル化する際，音声の特徴量と，内部状態を表す特徴量のベクトルが結合され，ひとつのベクトルとして扱われる．したがって，GMM

では 2 つの可視変数（音声と内部状態）の特徴量空間を明示的に分離することができない．また，2 つの可視変数の特徴量を結合したベクトルを用いて学習を行うため，特徴量空間の次元が大きくなり，モデルの表現能力次第では過学習の影響を受けやすいといえる．その後，GMM の代わりに多層のニューラルネットワークである Deep Neural Network（DNN）を用いて，音声の特徴量から HMM の内部状態を推定するハイブリッド方式 [2]（DNN-HMM）や Convolutional Neural Network（CNN）を用いた手法 [3] が提案され，GMM-HMM よりも高い性能を示すことが報告されている．

音声合成においても，音声認識と同様に HMM を用いた統計的手法 [4] が盛んに研究されている．この手法は音声のピッチや長さを同時にモデル化できる [5] などの理由から，波形接続型の音声合成 [6] よりも優れた点をもつ．しかし，HMM 音声合成では合成音声の質が問題となる [7]．その理由のひとつとして，決定木を用いてコンテキストクラスタリングを行う文脈依存 HMM では，テキストのコンテキストラベルと音声パラメータの依存関係を効率的にモデル化できないことがあげられる．そこで，Zen ら [8] は決定木を DNN で代替することで，合成音声の質を向上させる手法を提案した．

また，音声合成では近年，生の音声波形を，前処理を必要としないフレームワークを用いて直接モデル化する手法も提案されている [9]．この手法は，生の波形を用いて音声の時間的な依存性を表すことにより，単純な DNN を用いたフレーム独立な手法よりも合成音声の品質を改善することが報告されている．さらに，モデルの学習を高速化するために，いくつかの手法 [10, 11] が提案されている．しかしながら，これらのアプローチは統計的またはフレーム独立なアプローチと比較して，学習および合成をする際において依然として大きなコストを要する．

DNN に基づく手法は，GMM のようにテキストと音声といった 2 つの特徴量を 1 つのベクトルにまとめて学習を行うのではなく，それらを分離して扱う．そして，DNN を用いて，テキストから音声，あるいは音声からテキストへの単一方向の関係性を表現する．この手法は，非線形関数を多層に積重ねた DNN の表現能力が高いことや，DNN が入力層と出力層をもち，2 つの可視変数の特徴量空間が明示的に分離可能であることから，GMM を用いた手法よりも高い性能を示す．しかし，DNN は出力となる可視変数から，入力となる可視変数への逆向きの関係性を表現することができない．すなわち，音声とそれに対応するテキストには双方向に深い関係性があると考えられるが，DNN は合成器を構築する際に音声からテキスト

への逆向きの関係性を考慮していないので，構築した合成器のパラメータを認識器へと応用することができない．したがって，入力をテキスト，出力を音声として音声合成器を構築したのち，入力を音声，出力をテキストとする認識器を構築する場合には，再度学習を行わなければならない．これは認識器を構築した後で合成器を構築する場合も同じである．

バイナリ画像の認識，生成分野において，Deep Relational Model（DRM） [12] が画像・ラベル間の双方向の関係性を表現することで，画像の認識および生成を行うことができると報告されている．DRM は，DNN のように多層の隠れ層と，分離された 2 つの可視層からなる，エネルギー関数に基づく生成モデルである．しかし，従来の DRM は Bernoulli 分布で表現されるバイナリ値のみを扱うため，音声の特徴量を表現する正規分布や，テキストを表現するカテゴリカル分布を扱う音声認識および合成に応用することができない．そこで，本論文では，2 変数間の双方向変換を可能にする DRM を音声認識および合成へと応用するために，Gaussian-Categorical DRM（GCDRM）を定義し，テキスト・音声間の双方向の関係性を表現することで，DNN 音声認識器および DNN 音声合成器を同時に構築する手法を 1 つ目に提案する．そして，実験によって DRM を音声認識および音声合成に応用可能であることを示す．

また，音声信号処理の別の応用例として声質変換があげられる．声質変換とは，ソース話者の音声を，発話内容を保持しつつ変形することで，ターゲット話者の発話であるかのように認識させる技術である．この技術は喉頭摘出者の発声を補助したり，自分の歌声を所望の話者の歌声に変換するなど，物理的な制約を超えた音声コミュニケーションを実現することができる．

ソース話者とターゲット話者が同じ内容を発話した音声の対であるパラレルデータを使用する声質変換では，コードブックに基づく手法 [13] が提案されて以来，様々な手法が提案されている [14, 15, 16]．その中でも，GMM に基づく統計的手法 [17, 18] や DNN に基づく手法 [19, 20, 21] が盛んに研究されている．

GMM に基づく手法 [17, 18] では，ソース話者とターゲット話者の音響特徴量の結合分布をモデル化するために，最尤法基準で GMM のパラメータを推定する．そして，得られたパラメータをもとにソース話者の音声をターゲット話者の音声へと変換する．GMM に基づく変換法は，その柔軟性の高さから広く用いられている．

一方，DNN に基づく手法 [19, 20, 21] では，GMM のように話者ペアの音響特徴量の結合分布をモデル化するのではなく，DNN を用いて，ソース話者からター

ゲット話者の単一方向の関係性を表現する．これらの手法は，上で述べたようにDNN が高い表現能力や，入出力を明示的に分離できる構造をもつことから，GMMに基づく手法よりも性能を改善することが報告されている．

しかし，音声認識および合成の場合と同様に，DNN は一度の学習でソース話者からターゲット話者への単一方向の声質変換器しか構築することができない．そこで本論文では，DNN のような表現力と構造をもつ生成モデルである DRM を声質変換へと応用するために，Gaussian-Gaussian DRM（GGDRM）を定義し，ソース話者およびターゲット話者の結合分布をモデル化する手法を 2 つ目に提案する．また，実験によって，DRM を声質変換に応用可能であることを示す．

まとめれば，本論文では，DRM を音声信号処理へと応用する手法を提案する．まず 1 つ目に DRM を音声認識および音声合成へ応用する手法について述べる．次に，DRM を声質変換へ応用する手法について述べる．

1.2　本論文の構成

以下に本論文の構成をまとめる．

第 2 章では，音声信号処理の基礎知識について説明する．

第 3 章では，DNN の基礎知識と，DNN の音声信号処理への応用について説明する．

第 4 章では，エネルギー関数に基づく種々のモデルについて説明する．

第 5 章では，本論文が提案する DRM の音声認識・音声合成への応用方法について説明する．また，提案法の実験結果および考察について述べる．

第 6 章では，本論文が提案する DRM の声質変換への応用方法について説明する．また，提案法の実験結果および考察について述べる．

第 7 章では結論および本研究の今後の課題についてまとめる．

第2章

音声信号処理の先行研究

音声は言語情報を伝達する声である．音声には言語情報以外に，話者に関する様々な情報が含まれている．音声言語による人間・コンピュータ間のやりとりを可能にするヒューマンマシンインターフェースをはじめとする様々な音声言語システムを実現するためには，音声の認識や合成などを行わなければならない．そして，音声の認識や合成などを行うためには，正確かつ能率の良い音声のパラメータ表現を考える必要がある．本章では，まず音声信号処理においてよく用いられる音声のパラメータ表現について説明し，その後，音声信号処理の例として音声認識，音声合成および声質変換について説明する．

2.1　音声信号のパラメータ表現

音声の認識や合成，変換を行うためには，音声信号になんらかの処理を行わなければならない．音声の認識も合成も変換も，それらを実現するシステムの性能は人間の特性を基準に評価されるので，音声信号処理は音声に関する人間の特性を考慮して行う必要がある．すなわち，音声信号は，人間が発生した音声がもつ特徴をもつものとして処理する必要がある．

音声スペクトルの大まかな形を表す包絡線をスペクトル包絡と呼ぶ．人間の発生した音声のスペクトル包絡は，比較的次数の低い三角多項式によって表された対数スペクトルによってよく近似できることが知られている．したがって，音声信号のスペクトル包絡はケプストラム（cepstrum）を係数とする三角多項式による対数スペクトルの形で効率よく正確に表現されることになる．さらに，周波数分解能に対する聴覚の特性が，低周波域では高周波域に比べてかなり高く，その周

波数目盛はしばしば音高の知覚尺度であるメルスケールで表されるので，音声のスペクトルはメルスケールのような非直線周波数目盛上の次数の低い三角多項式による対数スペクトル，すなわちメルケプストラム（mel-cepstrum）による対数スペクトルによって精密に表されると考えられる．

すなわち，人間の発声器官と聴覚の特性を考慮すると，音声のスペクトル包絡はメルケプストラムを係数とする三角多項式によるメル対数スペクトルの形で表されることになる．そこで本節では，ケプストラム，メル周波数目盛およびメルケプストラムについて説明する [22, 23, 24].

2.1.1　ケプストラム

信号 $x[n]$ のフーリエ変換を $X(e^{j\Omega})$ としたとき，$x[n]$ のケプストラム $c[m]$ は

$$c[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|X(e^{j\Omega})|^2 e^{j\Omega m} d\Omega = \frac{1}{\pi}\int_{-\pi}^{\pi} \ln|X(e^{j\Omega})| e^{j\Omega m} d\Omega \tag{2.1}$$

で定義される．ここで，m はケフレンシ（quefrency）と呼ばれる時間量を表し，単位時間（信号の標本化間隔）の倍数である．

式 (2.1) の積分を台数公式で近似する．全積分区間の割合の数 N が十分大きければ，通常の信号のフーリエ変換の対数 2 乗振幅 $\ln|X(e^{j\Omega})|^2$ の値は隣り合った離散周波数間でほぼ直線的に変化するので，式 (2.1) は

$$c[m] = \frac{1}{N}\sum_{k=0}^{N-1} \ln|X[k]|^2 e^{j\frac{2\pi}{N}km} \tag{2.2}$$

と書き換えられる．ここで，$X[k]$ は

$$X[k] = X(e^{j\Omega})|_{\Omega=j2\pi k/N} = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \tag{2.3}$$

である．$X[k]$ の値は高速フーリエ変換を利用して効率良く求めることができるため，実際の信号からケプストラムを求める場合，式 (2.2) を用いる．

2.1.2　メル周波数目盛

1 次のオールパスフィルタの位相特性によってメルスケールを近似することを考える．メルスケールを正確に表すことができ，フィルタ構成上も都合のよい非直

線周波数目盛として，伝達関数 $\Phi(z,\alpha)$ および周波数特性 $\Phi(e^{j\Omega},\alpha)$ がそれぞれ

$$\Phi(z,\alpha) = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}} \tag{2.4}$$

$$\Phi(e^{j\Omega},\alpha) = \exp(-\Omega(\widetilde{\Omega},\alpha)) \quad (\Omega=\omega\Delta t) \tag{2.5}$$

で与えられるオールパスフィルタの位相特性

$$\widetilde{\Omega}(\Omega,\alpha) = \Omega + 2\tan^{-1}\frac{\alpha\sin\Omega}{1-\alpha\cos\Omega} \tag{2.6}$$

による非直線周波数目盛

$$\widetilde{\Omega} = \widetilde{\Omega}_\alpha = \widetilde{\Omega}(\Omega,\alpha) \tag{2.7}$$

を考えることができる．ここで，$\Omega=\omega\Delta t$ であり，ω は角周波数，Δt はフィルタの単位遅延時間である．また，$1/\Delta t = f_s$ はサンプリング周波数である．この非直線周波数目盛 $\widetilde{\Omega}$ は，Δt が $100\mu s$（f_s=10kHz）のとき，パラメータ α を 0.35 前後の値にすれば，メルスケールをよく近似できる．メルスケールをよく近似する場合の非直線周波数目盛をメル周波数目盛と呼ぶ．

2.1.3　メルケプストラム

信号 $x[n]$ について

$$c_\alpha[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi}\hat{g}_\alpha(\widetilde{\Omega})e^{j\widetilde{\Omega}m}d\Omega \tag{2.8}$$

を信号 $x[n]$ のメルケプストラムと呼ぶ．ここで，$\hat{g}_\alpha(\widetilde{\Omega})$ は $x[n]$ のメル周波数上の対数スペクトルであり，

$$\hat{g}_\alpha(\widetilde{\Omega}) = \sum_{m=-M}^{M} c_\alpha[m]e^{-j\widetilde{\Omega}m} = c_\alpha[0] + 2\sum_{m=1}^{M} c_\alpha[m]\cos(\widetilde{\Omega}m) \tag{2.9}$$

と三角多項式で表される．ここで M はメルケプストラムの次数である．

2.2　音声認識

音声言語を用いて人間・コンピュータ間のコミュニケーションを行うためには，コンピュータは人間の発した音声に対応するテキストを自動的に決定しなければならない．この操作を自動音声認識（automatic speech recognition）あるいは単に音声認識（speech recognition）という．

音声認識における音声の基本単位

音声を構成する最小の単位を音素（phoneme）という．音素の種類は言語によって異なるが，その数は日本語でも英語でも 40 前後である．日本語における音素は，ほぼローマ字 1 文字に相当する．例えば,「音声」は実際には「おんせい」ではなく「おんせえ」と発音されるため，音素記号列で表すと「o N s e e」となる．ここで，音素記号「N」は撥音「ん」に相当する音素記号である．任意の音声は音素の組合せによって表現される．音声認識の基本単位として音素を採用するか単語を採用するかの違いは，音声認識システムの構成上大きな違いとなる．

音声認識の定式化

音声認識は，与えられた音声 $\boldsymbol{O}$ に対して，任意の記号（音素，単語など）$\boldsymbol{W}$ から，$p(W|\boldsymbol{O})$ を最大にする記号 W_{max} を求める問題である．

$$\begin{aligned} W_{max} &= \arg\max_{W} p(W|\boldsymbol{O}) \\ &= \arg\max_{W} \frac{p(\boldsymbol{O}|W)p(W)}{p(\boldsymbol{O})} \\ &= \arg\max_{W} p(\boldsymbol{O}|W)p(W) \end{aligned} \tag{2.10}$$

と定式化される．ただし，W に関する最大化に対して，$p(\boldsymbol{O})$ は定数であることに注意する．式 (2.10) における $p(W)$ は言語モデルと呼ばれ，単語の部分列が出現する確率のモデルの積として与えられる．また，$p(\boldsymbol{O}|W)$ は音響モデルと呼ばれる．

2.3　音声合成

入力されたテキストに対応する音声を生成することを音声合成（speech synthesis）やテキスト読上げ（text-to-speech; TTS）という．また，これにより生成された音声を合成音声という．音声合成は，音声言語を用いた人間・コンピュータ間のコミュニケーションにおいて，コンピュータが人間に情報を伝達する際に用いられる．

音声合成の定式化

音声合成は，与えられた単語列 $\boldsymbol{W}$ に対して，任意の音声 $\boldsymbol{O}$ から，$p(O|\boldsymbol{W})$ を最大にする音声 O_{max} を求める問題である．すなわち，

$$\begin{aligned} O_{max} &= \arg\max_{O} p(O|\boldsymbol{W}) \\ &= \arg\max_{O} \frac{p(\boldsymbol{W}|O)p(O)}{p(\boldsymbol{W})} \\ &= \arg\max_{O} p(\boldsymbol{W}|O)p(O) \end{aligned} \tag{2.11}$$

と定式化される．ただし，O に関する最大化に対して，$p(\boldsymbol{W})$ は定数であることに注意する．音声認識の定式化 (2.10) と音声合成の定式化 (2.11) は，それぞれ対象的な関係となっている．

2.4　声質変換

声質変換とは，ソースとなる話者の発話を変形することで，発話内容を保持したまま，ターゲットとなる話者の発話であるかのように認識させる技術である．声質の変換処理は，ソース話者とターゲット話者が同じ内容の文を読上げた音声の対であるパラレルデータを用いて変換関数を構築することで実現される．

声質変換は，ソース話者の音声 $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$，$\boldsymbol{x}$ と時間フレームの対応付けがされたターゲット話者の音声 $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$，そして変換関数を $F(\cdot)$ として，

$$\boldsymbol{y} = F(\boldsymbol{x}) \tag{2.12}$$

と，回帰問題として定式化することができる．

それぞれの話者による発話間の時間フレームの対応付けは，Dynamic Programming（DP）マッチング [25] によって行われる．DP マッチングは動的計画法を用いた手法で，以下の手順で実行される．

1. 1 つ目のデータを $a_1, \ldots, a_N$，2 つ目のデータを $b_1, \ldots, a_M$ として，$N \times M$ のグリッドグラフを構成する．
2. 頂点 $(i,\ j)$ に対して，a_i と b_j の非類似度をコストとして設定する．
3. 頂点 $(1,\ 1)$ から頂点 $(N,\ M)$ までのコストが最小となるパスを検索する．
4. 最小コストの値からパスを逆順に求めていき，頂点 $(i,\ j)$ を通過するならば，a_i と b_j が対応している．

第3章

Deep Neural Networkと音声信号処理

Deep Neural Network（DNN）とは，多階層の隠れ層（hidden layer）をもつ階層型のニューラルネットワークである．本章では，DNNの概要 [26] について説明をしたのち，DNNの音声信号処理への応用について紹介する．

3.1 Deep Neural Network

3.1.1 DNNの構造

DNNの構造を図 3–1 に示す．DNNは，入力となる値を受取る入力層（input layer），入力層に近い層から順に値を計算する，L 層からなる隠れ層，そして隣接する隠れ層から計算される値を出力する出力層（output layer）で構成される．

図中の丸がひとつのニューロンユニットを表し，各層はいくつかのユニットで構成される．各ユニットは入力を受け取ると活性化関数（activation function）$f(\cdot)$ により出力を決定し，次の層へと値を出力する．活性化関数として，次式で定義されるシグモイド関数 $\sigma(\cdot)$ がよく用いられる．

$$\sigma(x) = \frac{1}{1+\exp(-x)} \tag{3.1}$$

各隠れ層の入力を $\boldsymbol{u}^{(l)}$，出力を $\boldsymbol{z}^{(l)}$ で表すと，入力 $\boldsymbol{x}$ が与えられたときの隠れ層（$l=1$）の入出力は次のように計算される．

$$\begin{aligned} \boldsymbol{u}^{(1)} &= \boldsymbol{W}^{(1)T}\boldsymbol{x} + \boldsymbol{b}^{(1)} \\ \boldsymbol{z}^{(1)} &= f(\boldsymbol{u}^{(1)}) \end{aligned}$$

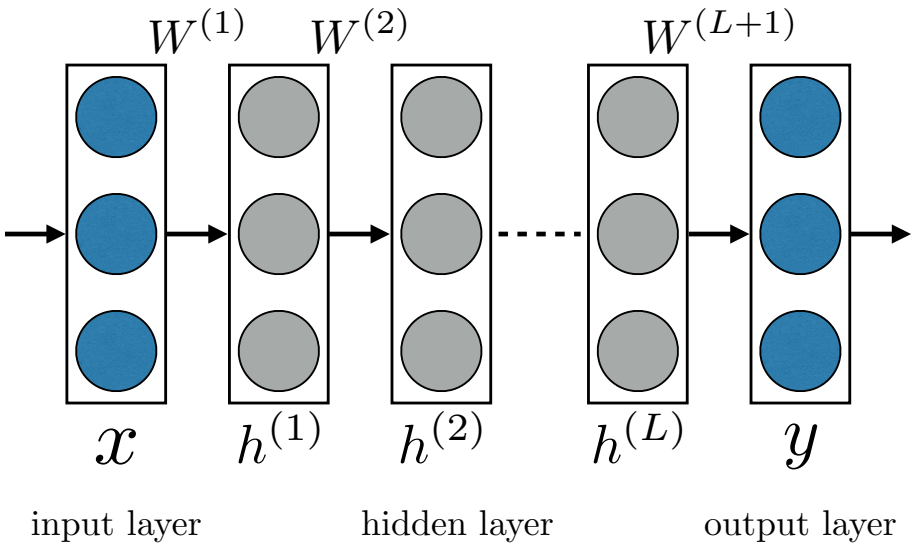


図 3–1: Graphical representation of a deep neural network.

ここで，$\boldsymbol{W}^{(1)}$ は入力層・隠れ層（$l = 1$）間の結合重みを表し，入力層のユニット数が I，1 番目の隠れ層（$l = 1$）のユニット数が J_1 のとき，$\boldsymbol{W}^{(1)} \in \mathbb{R}^{I \times J_1}$ である．また，$\boldsymbol{b}^{(1)}$ は隠れ層（$l = 1$）のバイアス項を表し，$\boldsymbol{b}^{(1)} \in \mathbb{R}^{J_1}$ である．

次に，隠れ層（$l = 2, \ldots, L$）の入出力は，第 l 番目の隠れ層の隠れ変数を $\boldsymbol{h}^{(l)}$ として，以下のように計算される．

$$\begin{aligned} \boldsymbol{u}^{(l)} &= \boldsymbol{W}^{(l)T} \boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)} \\ \boldsymbol{z}^{(l)} &= f(\boldsymbol{u}^{(l)}) \end{aligned}$$

ただし，$\boldsymbol{W}^{(l)}$ は隠れ層 $l-1$・隠れ層 l 間の結合重みを表し，隠れ層 $l-1$ のユニット数が J_{l-1}，隠れ層 l のユニット数が J_l のとき，$\boldsymbol{W}^{(l)} \in \mathbb{R}^{J_{l-1} \times J_l}$ である．また，$\boldsymbol{b}^{(l)}$ は隠れ層 l のバイアス項を表し，$\boldsymbol{b}^{(l)} \in \mathbb{R}^{J_l}$ である．

そして，出力層の入出力 $\boldsymbol{u}^{(L+1)}$, $\boldsymbol{z}^{(L+1)}$ は

$$\begin{aligned} \boldsymbol{u}^{(L+1)} &= \boldsymbol{W}^{(L+1)T} \boldsymbol{h}^{(L)} + \boldsymbol{b}^{(L+1)} \\ \boldsymbol{z}^{(L+1)} &= f(\boldsymbol{u}^{(L+1)}) \end{aligned}$$

と計算される．ここで，$\boldsymbol{W}^{(L+1)}$ は隠れ層 l・出力層間の結合重みを表し，出力層のユニット数が K のとき，$\boldsymbol{W}^{(L+1)} \in \mathbb{R}^{J_L \times K}$ である．また，$\boldsymbol{b}^{(L+1)}$ は出力層のバイアス項を表し，$\boldsymbol{b}^{(L+1)} \in \mathbb{R}^{K}$ である．そして，ネットワークの最終的な出力を $\boldsymbol{y} \equiv \boldsymbol{z}^{(L+1)}$ とする．

このように，DNN は与えられた入力 $\boldsymbol{x}$ に対して，入力層から出力層へと上の計算を繰り返し，値を順伝播させていくことで出力 $\boldsymbol{y}$ を計算する．DNN の全ての層のパラメータ $\boldsymbol{W}$ と $\boldsymbol{b}$ をまとめて $\boldsymbol{\theta}$ で表すと，DNN はパラメータ $\boldsymbol{\theta}$ もつ関数 $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}; \boldsymbol{\theta})$ と表現することができる．DNN はパラメータを変えることで，様々な関数を表現することができる．

3.1.2　出力層の設計と誤差関数

DNN が表現する関数 $\boldsymbol{y}(\boldsymbol{x}; \boldsymbol{\theta})$ は，パラメータ $\boldsymbol{\theta}$ を変えることで変化する．目標とする関数は，その具体的なパラメータはわからないが，関数の入力と出力のペアが複数与えられている．ある入力 $\boldsymbol{x}^t$ に対する望ましい出力を $\boldsymbol{d}^t$ とし，このような入出力のペアが複数 $\mathbf{D} = \{(\boldsymbol{x}^1, \boldsymbol{d}^1), \ldots, (\boldsymbol{x}^T, \boldsymbol{d}^T)\}$ のように与えられているとする．この集合 $\mathbf{D}$ を学習データ（training data）や訓練データと呼ぶ．

このとき，DNN のパラメータ $\boldsymbol{W}$ を調整することで，学習データに含まれる入出力のペアを再現することを考える．すなわち，学習データ $\mathbf{D}$ に含まれるすべての入出力ペア $(\boldsymbol{x}^t, \boldsymbol{d}^t)$ について，DNN の出力 $\boldsymbol{y}(\boldsymbol{x}^t; \boldsymbol{\theta})$ が $\boldsymbol{d}^t$ となるべく近くなるように $\boldsymbol{\theta}$ を調整する．この操作を学習と呼ぶ．学習の際に，DNN の出力 $\boldsymbol{y}(\boldsymbol{x}^t; \boldsymbol{\theta})$ と，正解となる学習データ $\boldsymbol{d}^t$ の近さを測る尺度のことを，誤差関数（error function）あるいは損失関数という．DNN の学習では，扱う問題の種類に応じて，誤差関数と出力層の活性化関数を適切に設定する必要がある．

まず，回帰のように出力として実数値をとる場合は，出力層の活性化関数は恒等関数となる．そして，誤差関数は二乗誤差

$$E(\boldsymbol{\theta}) = \frac{1}{2} \sum_{t=1}^{T} \|\boldsymbol{d}^t - \boldsymbol{y}(\boldsymbol{x}^t; \boldsymbol{\theta})\|_2^2 \tag{3.2}$$

を採用するのが一般的である．

また，入力 $\boldsymbol{x}$ を K 個のクラスに分類する問題である多クラス分類では，入力に対応するクラスラベルを，one-hot ベクトルを用いて表現する．したがって，出力層の k 番目のユニットの出力を

$$y_k \equiv z_k^{(L+1)} = \frac{\exp(u_k^{(L+1)})}{\sum_{j=1}^{K} \exp(u_j^{(L+1)})} \tag{3.3}$$

とするソフトマックス関数を用いる．ただし，$u_k^{(L+1)}$ は出力層の k 番目のユニット

が隠れ層 L から受取る入力である．そして，誤差関数は交差エントロピー誤差

$$E(\boldsymbol{\theta}) = -\sum_{t=1}^{T}\sum_{k=1}^{K} d_k^t \log y_k(\boldsymbol{x}^t;\boldsymbol{\theta}) \tag{3.4}$$

を用いる．

3.1.3　DNN の学習

DNN の学習では，誤差関数 E が最小となるようパラメータ $\boldsymbol{\theta}$ を更新する．パラメータの更新には勾配法が用いられる．誤差関数 E と $\boldsymbol{\theta}$ から計算される更新量を $\Delta\boldsymbol{\theta}$ とすると，パラメータの更新式は

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta} \tag{3.5}$$

となる．パラメータの更新量 $\Delta\boldsymbol{\theta}$ は，誤差関数 E の $\boldsymbol{\theta}$ に関する偏微分

$$\Delta\boldsymbol{\theta} = -\eta\frac{\partial E(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \tag{3.6}$$

で計算される．ここで，η は学習率（learning rate）と呼ばれる，小さな正の値である．

例えば，誤差関数が二乗誤差であるとき，第 $l-1$ 層の i 番目のユニットから第 l 層の j 番目のユニットへの結合重み $W_{ij}^{(l)}$ の更新量 $\Delta W_{ij}^{(l)}$ は次のように計算される．

$$\Delta W_{ij}^{(l)} = -\eta\frac{\partial E(\boldsymbol{\theta})}{\partial W_{ij}^{(l)}} = -\eta\delta_j^{(l)} z_i^{(l-1)} \tag{3.7}$$

ただし，$\delta_j^{(l)}$ は誤差信号と呼ばれ，合成関数の微分法に従い，出力層で計算される二乗誤差から計算される $\delta_j^{(L+1)}$ を用いて以下のように再帰的に計算される．

$$\delta_j^{(L+1)} = -(d_j - z_j^{(L+1)})z_j^{(L+1)}(1 - z_j^{(L+1)}) \tag{3.8}$$

$$\delta_j^{(l-1)} = \boldsymbol{W}_{i:}^{(l)}\boldsymbol{\delta}^{(l)} z_j^{(l-1)}(1 - z_j^{(l-1)}) \tag{3.9}$$

ここで，$\boldsymbol{W}_{i:}$ は $\boldsymbol{W}$ から i 行目を抜出したベクトルを表す．この計算過程が，出力層の誤差を入力層側へと逆向きに伝播させていく形になっていることから，この学習方法を Back Propagation（BP）法と呼ぶ．

3.2　DNN の音声信号処理への応用

本節では，DNN を音声認識，音声合成および声質変換へと応用する方法について説明する．

DNN を用いた音声認識

DNN を音声認識へ応用する場合，音響特徴量（メルケプストラムおよびその動的特徴ベクトル [27]）を入力 $\boldsymbol{x}$ に，音素を出力 $\boldsymbol{y}$ に割当てる．動的特徴ベクトルとは，メルケプストラムのような静的な特徴量ベクトル $\boldsymbol{c}_t$ の系列に関する時間微分を表し，

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau)\boldsymbol{c}_{t+\tau} \tag{3.10}$$

$$\Delta^2 \boldsymbol{c}_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} w^{(2)}(\tau)\Delta\boldsymbol{c}_{t+\tau} \tag{3.11}$$

により計算される．ここで，$w^{(1)}(\tau)$, $w^{(2)}(\tau)$ はそれぞれ 1 次，2 次の動的特徴を求めるための重み係数である．また，音素はベクトルの one-hot 表現により表される．

ここでの目的は，式 (2.10) における音響モデル $p(\boldsymbol{O}|W)$ を DNN を用いて表現することである．ここで，$p(\boldsymbol{O}|W)$ は音響特徴量から音素を推定する問題 $p(W|\boldsymbol{O})$ の逆問題となっている．学習はこのような推定問題の逆問題となっており，仮定したモデル（音声認識器）のパラメータを，観測データから推定する問題になっている．

音声認識は，与えられた音響特徴量に対応する音素を当てる多クラス分類の問題であるから，出力層の活性化関数にはソフトマックス関数を，誤差関数にはクロスエントロピー誤差を用いる．

DNN を用いた音声合成

DNN を音声合成へ応用する場合，言語特徴量を入力 $\boldsymbol{x}$ に，音響特徴量を出力 $\boldsymbol{y}$ に割当てる．言語特徴量には，当該フレームの音素およびその前後 2，3 フレームの音素と，テキストのコンテキストラベルが含まれる．コンテキストラベルとは，テキストを構文解析して得られるいくつかの言語学的な情報を表し，当該フレームが含まれる単語の品詞 ID，活用形やアクセントの位置などから構成される．

ここでの目的は，式 (2.11) における $p(\boldsymbol{W}|O)$ を DNN を用いて表現することである．ここで，$p(\boldsymbol{W}|O)$ は音声認識のときと同様に，言語特徴量から音響特徴量を生成する問題 $p(O|\boldsymbol{W})$ の逆問題となっている．

音声合成は，与えられた言語特徴量に対応する音響特徴量を出力する回帰問題であるから，出力層の活性化関数には恒等関数を，誤差関数には二乗誤差を用いる．

DNN を用いた声質変換

DNN を音声合成へ応用する場合，ソース話者の音響特徴量を入力 $\boldsymbol{x}$ に，ターゲット話者の音響特徴量を出力 $\boldsymbol{y}$ に割当てる．パラレルデータを用いた学習では，ソース話者の音声とターゲット話者の音声のフレーム長が異なるため，DP マッチングを用いてフレームの対応付けを行なったのち，学習を行う．

声質変換は，ソース話者の音響特徴量に対応する，ターゲット話者の音響特徴量を出力する回帰問題であるから，出力層の活性化関数には恒等関数を，誤差関数には二乗誤差を用いる．

第4章

Energy-Based Model

本章では，生成モデルに対するアプローチのひとつである，エネルギー関数（energy function）に基づくモデルについて説明する．

未知の確率分布 $p_g(\boldsymbol{v})$ からそれぞれ独立に生成された複数の観測データ $\boldsymbol{v} = \{\boldsymbol{v}^1, \ldots, \boldsymbol{v}^T\}$ が与えられたとする．もし分布の詳細がわかれば，観測データの生成メカニズムも掌握できることになる．そこで，この未知の分布を，観測データを利用して再現することを考える．まず適当なパラメータ $\boldsymbol{\theta}$ をもつ生成モデル $p(\boldsymbol{v};\boldsymbol{\theta})$ を仮定する．そして，観測データを用いて $\boldsymbol{\theta}$ の値を変えることでモデルを調整し，分布 $p_g(\boldsymbol{v})$ の再現を試みる．パラメータ $\boldsymbol{\theta}$ は一般に最尤推定によって求められる．すなわち，複数の観測データがそれぞれ独立に生成されたとするとき，それらの結合確率を $\boldsymbol{\theta}$ の関数として捉えた尤度関数あるいはその対数をとった対数尤度関数 $\mathcal{L}(\boldsymbol{\theta}) = \log \prod_t p(\boldsymbol{v}^t;\boldsymbol{\theta})$ を考え，これを最大化する $\boldsymbol{\theta}$ をパラメータの推定値とする [28].

エネルギー関数に基づくモデルは，エネルギー関数を E とすれば

$$p(\boldsymbol{v};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\{-E(\boldsymbol{v};\boldsymbol{\theta})\} \tag{4.1}$$

と，確率分布として定義される [29]．式 (4.1) はボルツマン分布（Boltzmann distribution）あるいはギブス分布（Gibbs distribution）と呼ばれる．ここで，$Z(\boldsymbol{\theta})$ は $\boldsymbol{v}$ のすべての実現値に関する確率の総和 $\sum_{\boldsymbol{v}} p(\boldsymbol{v})$ を 1 にするための分配関数であり，

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{v}} \exp\{-E(\boldsymbol{v};\boldsymbol{\theta})\} \tag{4.2}$$

と表される．エネルギー関数に基づくモデルでは，エネルギー関数の値をより小さくするパラメータ $\boldsymbol{\theta}$ が，エネルギーを安定させるより良いパラメータであるとして，パラメータを調整する．エネルギー関数は観測データによって適当に設定される．

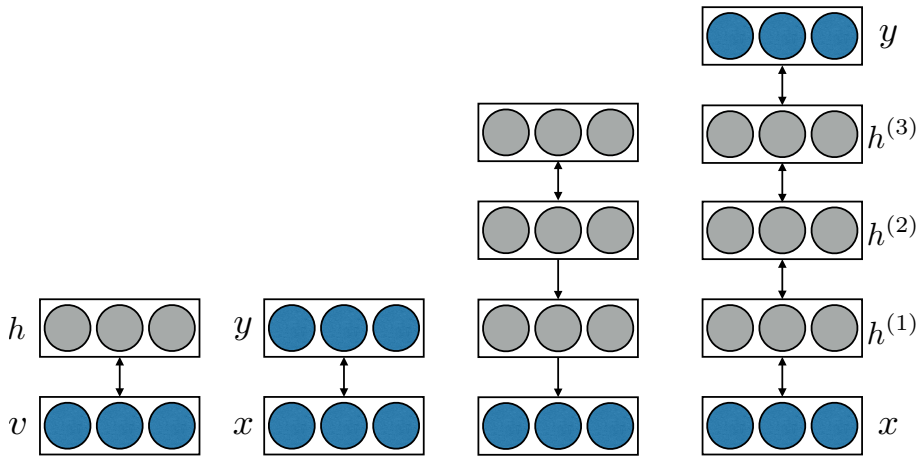


図 4–1: Graphical representations of (a) a restricted Boltzmann machine, (b) a bidirectional associative memory, (c) a deep belief network, and (d) a deep relational model.

4.1　Restricted Boltzmann Machine

本節では，エネルギー関数に基づく生成モデルである Restricted Boltzmann Machine（RBM） [30, 31]（図 4–1（a））について説明する．

Bernoulli-Bernoulli RBM

まず，可視変数 $\boldsymbol{v}$ がバイナリ値のみから構成される Bernoulli-Bernoulli RBM（BBRBM）について説明する．BBRBM は 2 層構造であり，片方の層は可視変数 $\boldsymbol{v} \in \{0, 1\}^I$ のみで構成される層（可視層），もう片方の層は隠れ変数 $\boldsymbol{h} \in \{0, 1\}^J$ のみで構成される層（隠れ層）である．BBRBM は完全二部グラフ上に定義される．すなわち，各ユニットは別の層のユニットとのみ結合をもち，同じ層のユニットとは結合をもたない．BBRBM の定義を次に示す．

$$p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\{-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})\} \tag{4.3}$$

ここで，E は BBRBM のエネルギー関数であり，

$$E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = -\boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{c}^T\boldsymbol{h} - \boldsymbol{v}^T\boldsymbol{W}\boldsymbol{h} \tag{4.4}$$

で定義される．ただし，$\boldsymbol{b} \in \mathbb{R}^I$, $\boldsymbol{c} \in \mathbb{R}^J$ はそれぞれ可視層，隠れ層のバイアス項を，$\boldsymbol{W} \in \mathbb{R}^{I \times J}$ は可視層・隠れ層間の結合重みを表し，学習により推定されるパラメー

タである．また，$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{v},\boldsymbol{h}} \exp\{-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})\}$ は分配関数である．

式 (4.3) の定義より，BBRBM の可視層および隠れ層の条件付き確率分布は以下で与えられる．

$$p(x_i = 1 \mid \boldsymbol{h}) = \sigma(b_i + \boldsymbol{W}_{i:}\boldsymbol{h}) \tag{4.5}$$

$$p(h_j = 1 \mid \boldsymbol{v}) = \sigma(c_j + \boldsymbol{W}_{:j}^T\boldsymbol{v}) \tag{4.6}$$

ただし，$\sigma(\cdot)$ はシグモイド関数である．また，式中の $\boldsymbol{W}_{i:}$, $\boldsymbol{W}_{:j}$ はそれぞれ $\boldsymbol{W}$ から i 行目を抜出したベクトル，j 列目を抜出したベクトルを表す．

BBRBM のパラメータ $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}\}$ は，$p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ の対数尤度 $\mathcal{L} = \log \prod_t p(\boldsymbol{x}^t, \boldsymbol{y}^t; \boldsymbol{\theta})$ が最大となるように推定される．対数尤度の各パラメータに関する偏微分は

$$\frac{\partial \mathcal{L}}{\partial b_i} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model} \tag{4.7}$$

$$\frac{\partial \mathcal{L}}{\partial c_j} = \langle h_j \rangle_{data} - \langle h_j \rangle_{model} \tag{4.8}$$

$$\frac{\partial \mathcal{L}}{\partial W_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \tag{4.9}$$

と計算される．ここで，$\langle\cdot\rangle_{data}$, $\langle\cdot\rangle_{model}$ はそれぞれデータの期待値，モデルの期待値を表す．前者は観測データの平均を計算することで得られるが，後者の計算は組み合わせ爆発の問題が生じる．そこで，学習によりパラメータを推定する際は，Contrastive Divergence（CD）法 [35] と呼ばれる手法が広く用いられている．CD 法では，RBM の分布から k 回サンプルしモデルの期待値を近似することで，計算コストを削減する．k は通常 1 に設定される．

Gaussian-Bernoulli RBM

次に，可視変数 $\boldsymbol{v}$ が実数値のみから構成される Gaussian-Bernoulli RBM（GBRBM）[32, 33] について説明する．GBRBM は BBRBM 同様，可視層と隠れ層からなる 2 層構造であり，隠れ変数がバイナリ値のみから構成される．しかし，BBRBM の可視変数がバイナリ値のみで構成されていたのに対し，GBRBM では，可視変数が実数値のみで構成される（$\boldsymbol{v} \in \mathbb{R}^I$）．GBRBM のエネルギー関数 E は

$$E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{1}{2}\left(\frac{\boldsymbol{v} - \boldsymbol{b}}{\boldsymbol{\sigma}}\right)^T\left(\frac{\boldsymbol{v} - \boldsymbol{b}}{\boldsymbol{\sigma}}\right) - \boldsymbol{c}^T\boldsymbol{h} - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}}\right)^T\boldsymbol{W}\boldsymbol{h} \tag{4.10}$$

で定義される．ただし，$\boldsymbol{\sigma}$ は可視変数の偏差を表す．また，式中の除算は要素ごとの除算を表す．

GBRBM はエネルギー関数に基づくモデルを用いて実数値を扱う手法として提案されている．しかし，この手法では分散項の影響で学習が不安定になるという問題がある．そこで，GBRBM の学習を改善するため，Improved GBRBM（IGBRBM）[34] が提案されている．IGBRBM のエネルギー関数は

$$E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{1}{2}\left(\frac{\boldsymbol{v}-\boldsymbol{b}}{\boldsymbol{\sigma}}\right)^T\left(\frac{\boldsymbol{v}-\boldsymbol{b}}{\boldsymbol{\sigma}}\right) - \boldsymbol{c}^T\boldsymbol{h} - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}\circ\boldsymbol{\sigma}}\right)^T\boldsymbol{W}\boldsymbol{h} \tag{4.11}$$

で定義される．ただし，$\circ$ はアダマール積を表す．

IGBRBM の可視層および隠れ層の条件付き確率分布は

$$p(x_i = v \mid \boldsymbol{h}) = \mathcal{N}\left(v \mid b_i + \boldsymbol{W}_{i:}\boldsymbol{h}, \sigma_i^2\right) \tag{4.12}$$

$$p(h_j = 1 \mid \boldsymbol{v}) = \sigma\left(c_j + \boldsymbol{W}_{:j}^T\left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}\circ\boldsymbol{\sigma}}\right)\right) \tag{4.13}$$

と計算される．ここで，$\mathcal{N}(\cdot \mid \mu, \sigma^2)$ は平均 μ，分散 σ^2 の正規分布を表す．

IGBRBM のパラメータ $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\sigma}\}$ は，$p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ の対数尤度 $\mathcal{L} = \log\prod_t p(\boldsymbol{x}^t, \boldsymbol{y}^t; \boldsymbol{\theta})$ が最大となるように推定される．対数尤度の各パラメータに関する偏微分は

$$\frac{\partial\mathcal{L}}{\partial b_i} = \left\langle\frac{1}{\sigma_i^2}v_i\right\rangle_{data} - \left\langle\frac{1}{\sigma_i^2}v_i\right\rangle_{model} \tag{4.14}$$

$$\frac{\partial\mathcal{L}}{\partial c_j} = \langle h_j\rangle_{data} - \langle h_j\rangle_{model} \tag{4.15}$$

$$\frac{\partial\mathcal{L}}{\partial W_{ij}} = \left\langle\frac{1}{\sigma_i^2}v_i h_j\right\rangle_{data} - \left\langle\frac{1}{\sigma_i^2}v_i h_j\right\rangle_{model} \tag{4.16}$$

と計算される．そして，偏差パラメータ σ_i を更新する際は，常に非負となるように，$z_i = \log\sigma_i^2$ を更新する．z_i に関する勾配は

$$\frac{\partial\mathcal{L}}{\partial z_i} = e^{-z_i}\left(\left\langle\frac{1}{2}(v_i - b_i)^2 - v_i\boldsymbol{W}_{i:}\boldsymbol{h}\right\rangle_{data} - \left\langle\frac{1}{2}(v_i - b_i)^2 - v_i\boldsymbol{W}_{i:}\boldsymbol{h}\right\rangle_{model}\right) \tag{4.17}$$

と計算される．IGBRBM の学習の際，$\langle\cdot\rangle_{model}$ の計算は BBRBM の学習と同様に，CD 法を用いて近似される．

4.2　Bidirectional Associative Memory

本節では，Kosko の Bidirectional Associative Memory（BAM）[36]（図 4–1（b））について説明する．

BAM はエネルギー関数に基づいて，ある可視変数 $\boldsymbol{x} \in \{0, 1\}^I$ と別の可視変数 $\boldsymbol{y} \in \{0, 1\}^K$ との関係性を表現するモデルである．BAM は RBM とは異なり，可視

層を 2 つもち，また隠れ層をもたない．BAM のエネルギー関数は次のように定義される．

$$E(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = -\boldsymbol{b}^T \boldsymbol{x} - \boldsymbol{d}^T \boldsymbol{y} - \boldsymbol{x}^T \boldsymbol{W} \boldsymbol{y} \tag{4.18}$$

ここで，$\boldsymbol{W} \in \mathbb{R}^{I \times K}$ は可視層間の結合重みを表す．また，$\boldsymbol{b} \in \mathbb{R}^I$, $\boldsymbol{d} \in \mathbb{R}^K$ はそれぞれ可視変数 $\boldsymbol{x}$, $\boldsymbol{y}$ のバイアス項である．Chen ら [37] は，BAM を確率密度関数として解釈し，BAM の結合確率分布を

$$P(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\{-E(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})\} \tag{4.19}$$

としている．ただし $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}, \boldsymbol{y}} \exp\{-E(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})\}$ は分配関数である．BAM のパラメータ $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{d}\}$ は RBM と同様，CD 法を用いた学習により推定される．

4.3　Deep Belief Network

本節では，RBM を多層に積み重ねた Deep Belief Network（DBN） [32, 38]（図 4–1（c））について説明する．

DBN は，与えられた学習データの深い階層表現を抽出することを期待し，RBM を多層に積み重ねて，逐次貪欲的に学習していくことで形成される．DBN は可視変数 $\boldsymbol{x} \in \{0, 1\}^I$ と L 層の隠れ変数 $\boldsymbol{h}^{(l)} \in \{0, 1\}^{J_l}$ $(l = 1, \ldots, L)$ より与えられる結合確率分布として，次のように定義される．

$$p(\boldsymbol{x}, \forall \boldsymbol{h}^{(l)}; \boldsymbol{\theta}) = \left(\prod_{k=0}^{L-2} p(\boldsymbol{h}^{(k)} | \boldsymbol{h}^{(k+1)}; \boldsymbol{\theta}^{(k+1)}) \right) p(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)}; \boldsymbol{\theta}^{(L)}) \tag{4.20}$$

ここで，可視層を第 0 層であるとみなし，$\boldsymbol{h}^{(0)} = \boldsymbol{x}$ としている．また，簡単のため，第 $l-1$ 層・第 l 層間の結合重み $\boldsymbol{W}^{(l)} \in \mathbb{R}^{J_{l-1} \times J_l}$ と第 l 層，第 $l-1$ 層のバイアス項 $\boldsymbol{b}^{(l)} \in \mathbb{R}^{J_l}$, $\boldsymbol{b}^{(l-1)} \in \mathbb{R}^{J_{l-1}}$ をまとめて $\boldsymbol{\theta}^{(l)}$ で表している（$J_0 = I$）．最上段の 2 層に関する結合確率分布 $p(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)}; \boldsymbol{\theta}^{(L)})$ は分配関数 Z を用いて

$$p(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)}; \boldsymbol{\theta}^{(L)}) = \frac{1}{Z(\boldsymbol{\theta}^{(L)})} \exp\{-E(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)}; \boldsymbol{\theta}^{(L)})\} \tag{4.21}$$

で表される RBM として定義されている．ここで，エネルギー関数 E は

$$E(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)}; \boldsymbol{\theta}^{(L)}) = -\boldsymbol{b}^{(L-1)T} \boldsymbol{h}^{(L-1)} - \boldsymbol{b}^{(L)T} \boldsymbol{h}^{(L)} - \boldsymbol{h}^{(L-1)T} \boldsymbol{W}^{(L)} \boldsymbol{h}^{(L)} \tag{4.22}$$

である．また，その他の層間の条件付き確率分布は

$$p(h_j^{(l)} = 1 | \boldsymbol{h}^{(l+1)}) = \sigma(b_j^{(l)} + \boldsymbol{W}_{j:}^{(l+1)} \boldsymbol{h}^{(l+1)}) \tag{4.23}$$

と計算される．

DBN の学習は，DBN 全体をまとめて学習するのではなく，2 層ごとに順に RBM とみなして学習する．最上段の 2 層以外の結合は上の層から下の層への有向となっているが，学習の際にはそれらを双方向の結合として扱い，学習の結果をあらためて DBN の有向リンクのパラメータとして割り当てる．DBN の学習の手順をまとめると，次のようになる．

1. 可視変数を $\boldsymbol{x} = \boldsymbol{h}^{(0)}$ とみなし，最下層の 2 層（可視層（$l = 0$）および隠れ層（$l = 1$））からなる RBM$^{(1)}$ を構成し，学習を行う．

2. 最後に学習した RBM$^{(l)}$ から $p(\boldsymbol{h}^{(l)} = 1|\boldsymbol{v} = \boldsymbol{h}^{(l-1)}) = \prod_{j \in \boldsymbol{h}^{(l)}} \sigma(b_j^{(l)} + \boldsymbol{W}_{:j}^{(l)T} \boldsymbol{h}^{(l-1)})$ を用いて値をサンプルする．

3. 隠れ層 l と隠れ層 $l + 1$ からなる RBM$^{(l)}$ を構成し，2. でサンプルした値を隠れ層 l に与えられる擬似的な観測データとみなして学習を行う．

4. 2. および 3. を層の数だけ繰り返す．

DBN を用いて教師あり学習を行う場合は，DBN の最上位の隠れ層の上に可視層をひとつ追加する．最上位の隠れ層と新たに追加した可視層との間のパラメータは，Support Vector Macine（SVM）などを用いて決定する [39]．あるいは，可視層を追加した DBN 全体を DNN とみなし，BP 法によりパラメータを調整する．これを fine-tuning [40] と呼ぶ．

4.4　Deep Relational Model

本節では，中鹿らの Deep Relational Model（DRM） [12]（図 4–1（d））について説明する．

DRM は RBM 同様，エネルギー関数に基づくモデルである．しかし，RBM が可視層をひとつしかもたないのに対し，DRM は 2 つの可視層をもつ．したがって，DRM は 2 つの可視変数について，それらを明示的に分離することができる構造をもつ．DRM はある可視変数 $\boldsymbol{x} \in \{0, 1\}^I$ と別の可視変数 $\boldsymbol{y} \in \{0, 1\}^K$，そして隠れ変数 $\boldsymbol{h}^{(l)} \in \{0, 1\}^{J_l} (l = 1, ..., L)$ より与えられる結合確率分布で定義される．さらに，DRM はこれまで説明したエネルギー関数に基づくモデルと同様，各ユニットは隣接する層のユニットのみと結合をもち，同じ層のユニットとは結合をもたな

い．DRM の結合確率分布は以下のように表される．

$$p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \sum_{\forall \boldsymbol{h}^{(l)}} p(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \boldsymbol{\theta}) \tag{4.24}$$

$$p(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\{-E(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \boldsymbol{\theta})\} \tag{4.25}$$

ここで，$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}} \exp\{-E(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \boldsymbol{\theta})\}$ は分配関数である．また，E は DRM のエネルギー関数であり

$$E(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \boldsymbol{\theta}) = -\boldsymbol{b}^T \boldsymbol{x} - \sum_{l=1}^{L} \boldsymbol{c}^{(l)T} \boldsymbol{h}^{(l)} - \boldsymbol{d}^T \boldsymbol{y} - \boldsymbol{x}^T \boldsymbol{W}^{(1)} \boldsymbol{h}^{(1)} - \sum_{l=2}^{L} \boldsymbol{h}^{(l-1)T} \boldsymbol{W}^{(l)} \boldsymbol{h}^{(l)} - \boldsymbol{h}^{(L)T} \boldsymbol{W}^{(L+1)} \boldsymbol{y} \tag{4.26}$$

で定義される．ただし，$\boldsymbol{b} \in \mathbb{R}^I$，$\boldsymbol{c}^{(l)} \in \mathbb{R}^{J_l}$ および $\boldsymbol{d} \in \mathbb{R}^K$ はそれぞれ 1 つ目の可視層，隠れ層 l および 2 つ目の可視層のバイアス項を，$\boldsymbol{W}^{(1)} \in \mathbb{R}^{I \times J_1}$，$\boldsymbol{W}^{(l)} \in \mathbb{R}^{J_{l-1} \times J_l}$ および $\boldsymbol{W}^{(L+1)} \in \mathbb{R}^{J_L \times K}$ はそれぞれ 1 つ目の可視層・隠れ層（$l = 1$）間，隠れ層 $l-1$・隠れ層 l 間および隠れ層 L・2 つ目の可視層間の結合重みを表し，いずれも学習により推定されるパラメータである．

式 (4.24) および (4.25) の定義より，各層の条件付き確率分布は以下で与えられる．

$$p(x_i = 1 \mid \boldsymbol{h}^{(1)}) = \sigma(b_i + \boldsymbol{W}^{(1)}_{i:} \boldsymbol{h}^{(1)}) \tag{4.27}$$

$$p(h^{(l)}_j = 1 \mid \boldsymbol{h}^{(l-1)}, \boldsymbol{h}^{(l+1)}) = \sigma(c^{(l)}_j + \boldsymbol{W}^{(l)T}_{:j} \boldsymbol{h}^{(l-1)} + \boldsymbol{W}^{(l+1)}_{j:} \boldsymbol{h}^{(l+1)}) \tag{4.28}$$

$$p(y_k = 1 \mid \boldsymbol{h}^{(L)}) = \sigma(d_k + \boldsymbol{W}^{(L)T}_{:k} \boldsymbol{h}^{(L)}) \tag{4.29}$$

ただし，式 (4.28) において $\boldsymbol{h}^{(0)} = \boldsymbol{x}, \boldsymbol{h}^{(L+1)} = \boldsymbol{y}$ としている．

DRM のパラメータ $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}^{(l)}, \boldsymbol{d}\}$ は，$p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ の対数尤度 $\mathcal{L} = \log \prod_t p(\boldsymbol{x}^t, \boldsymbol{y}^t; \boldsymbol{\theta})$ が最大となるように推定される．対数尤度の各パラメータに関する偏微分は

$$\frac{\partial \mathcal{L}}{\partial b_i} = \langle x_i \rangle_{data} - \langle x_i \rangle_{model} \tag{4.30}$$

$$\frac{\partial \mathcal{L}}{\partial c^{(l)}_j} = \langle h^{(l)}_j \rangle_{data} - \langle h^{(l)}_j \rangle_{model} \tag{4.31}$$

$$\frac{\partial \mathcal{L}}{\partial d_k} = \langle y_k \rangle_{data} - \langle y_k \rangle_{model} \tag{4.32}$$

$$\frac{\partial \mathcal{L}}{\partial W^{(l)}_{ij}} = \begin{cases} \langle x_i h^{(1)}_j \rangle_{data} - \langle x_i h^{(1)}_j \rangle_{model} & (l = 1) \\ \langle h^{(l-1)}_i h^{(l)}_j \rangle_{data} - \langle h^{(l-1)}_i h^{(l)}_j \rangle_{model} & (l = 2, \ldots, L) \\ \langle h^{(L)}_i y_j \rangle_{data} - \langle h^{(L)}_i y_j \rangle_{model} & (l = L+1) \end{cases} \tag{4.33}$$

と計算される．ここで，可視変数 $\boldsymbol{x}$, $\boldsymbol{y}$ のデータの期待値 $\langle \boldsymbol{x} \rangle_{data}$, $\langle \boldsymbol{y} \rangle_{data}$ は観測デー

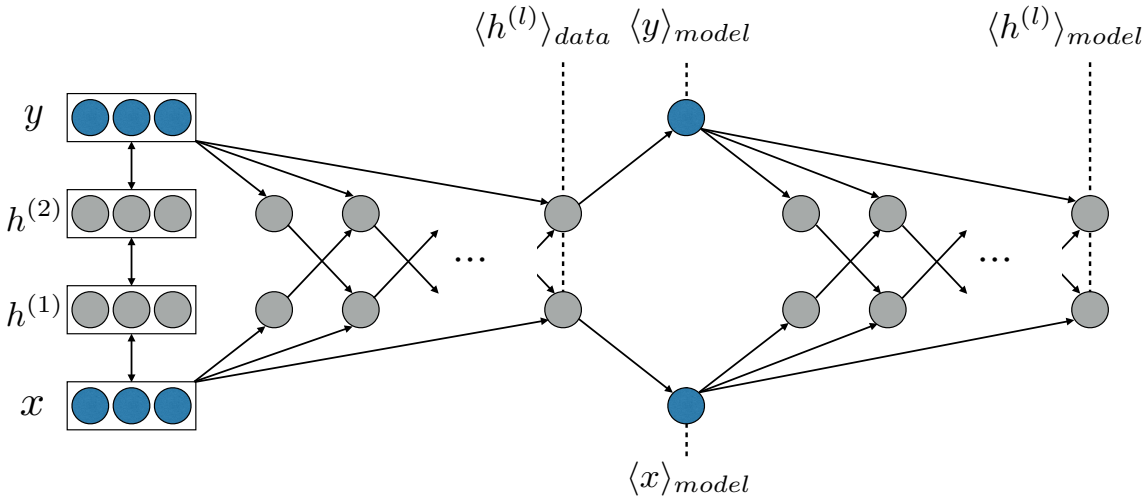


図 4–2: Calculating expectations using mean field update in the training of a DRM.

タの平均を計算することで得られる．また，隠れ変数 $\boldsymbol{h}^{(l)}$ のデータの期待値 $\langle \boldsymbol{h}^{(l)} \rangle_{data}$ は図 4–2 に示すように，$\boldsymbol{x}$, $\boldsymbol{y}$ に観測データを与えて，式 (4.28) より隠れ層の値を T 回更新することで得られる．一方，モデルの期待値 $\langle \cdot \rangle_{model}$ の計算は組合せ爆発の問題が生じる．そこで，DRM の学習では，平均場近似を用いてモデルの期待値を計算する．まず，可視変数に関するモデルの期待値は，図 4–2 のように，$\langle \boldsymbol{h}^{(l)} \rangle_{data}$ を用いて式 (4.27) および式 (4.29) から計算する．そして，得られた $\langle \boldsymbol{x} \rangle_{model}$, $\langle \boldsymbol{y} \rangle_{model}$ を用いて式 (4.28) より，隠れ層の値を T 回更新することで $\langle \boldsymbol{h}^{(l)} \rangle_{model}$ を得る．

DRM の学習の前に，RBM を用いて事前学習を行う．DRM の事前学習では，DBN の学習のように，まず可視層と隠れ層からなる RBM を最初に構築し学習する．次に，学習した RBM からサンプルした隠れ層の値を擬似的な観測データとして，隠れ層間のパラメータを学習する．ただし，DBN と違って，DRM は可視層を 2 つもつため，上下の可視層から内側へと順に RBM を構築し学習を行う．

第5章

DRMの音声認識・合成への応用

2.2節および2.3節で述べたとおり，音声認識および音声合成はそれぞれ式(2.10), (2.11)で定式化される．ここで，記号$\boldsymbol{W}$と音声$\boldsymbol{O}$の結合確率はベイズの定理より

$$p(\boldsymbol{W}, \boldsymbol{O}) = p(\boldsymbol{O}|\boldsymbol{W})p(\boldsymbol{W}) \tag{5.1}$$

$$= p(\boldsymbol{W}|\boldsymbol{O})p(\boldsymbol{O}) \tag{5.2}$$

と書ける．式(5.1)，(5.2)はそれぞれ式(2.10)，(2.11)と対応していることがわかる．したがって，$\boldsymbol{W}$と$\boldsymbol{O}$の結合確率を表現することができれば，音声認識器と音声合成器を同時に構築し，音声・テキスト間の双方向変換を行うことができる．そこで，4.4節で説明したDRMを用いて，$p(\boldsymbol{W}, \boldsymbol{O})$を表現する手法を提案する．

テキストを解析して得られる言語特徴量は，音素を表すバイナリ値と，コンテキストラベルを表す実数値からなる．また，音声パラメータを生成するために必要な音響特徴量は実数値からなる．したがって，テキスト・音声間の双方向変換を実現し，音声認識器と音声合成器を同時に構築するためには，4.4節で説明した従来のDRMのようにBernoulli分布ではなく，音素を表すカテゴリカル分布と，テキストのコンテキストラベルおよび音響特徴量を表す正規分布を同時に扱う必要がある．そこで，本章では従来のDRMをカテゴリカル分布および正規分布へと拡張したGaussian-Categorical DRM（GCDRM）を定義する．また，GCDRMを用いた音声認識および音声合成手法について述べる．そして，実験により，GCDRMを用いて音声認識および音声合成を行うことが可能であることを示す．

5.1 Gaussian-Categorical DRM

5.1.1 GCDRM の定義

4.1 節で述べたように，分散項を考慮して実数値を扱うエネルギー関数に基づくモデルとして IGBRBM が提案されている．また，音声合成のための DNN を事前学習する手法として Mixed GBRBM，Mixed Categorical-Bernoulli RBM（Mixed CBRBM）[41] が提案されている．IGBRBM，Mixed GBRBM および Mixed CBRBM のエネルギー関数を参考に，GCDRM のエネルギー関数を次のように定義する．

$$
\begin{aligned}
E(\boldsymbol{x}^c, \boldsymbol{x}^g, \boldsymbol{y}^c, \boldsymbol{y}^g, \forall \boldsymbol{h}^{(l)}; \boldsymbol{\theta}) \quad = \\
&\frac{1}{2}\left(\frac{\boldsymbol{x}^g - \boldsymbol{b}^g}{\boldsymbol{\sigma}^{(x)g}}\right)^T\left(\frac{\boldsymbol{x}^g - \boldsymbol{b}^g}{\boldsymbol{\sigma}^{(x)g}}\right) - \left(\frac{\boldsymbol{x}^g}{\boldsymbol{\sigma}^{(x)g} \circ \boldsymbol{\sigma}^{(x)g}}\right)^T \boldsymbol{W}^{(1)g}\boldsymbol{h}^{(1)} - \boldsymbol{b}^{cT}\boldsymbol{x}^c - \boldsymbol{x}^{cT}\boldsymbol{W}^{(1)c}\boldsymbol{h}^{(1)} \\
&- \sum_{l=1}^{L} \boldsymbol{c}^{(l)T}\boldsymbol{h}^{(l)} - \sum_{l=2}^{L} \boldsymbol{h}^{(l-1)T}\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l)} \\
&+ \frac{1}{2}\left(\frac{\boldsymbol{y}^g - \boldsymbol{d}^g}{\boldsymbol{\sigma}^{(y)g}}\right)^T\left(\frac{\boldsymbol{y}^g - \boldsymbol{d}^g}{\boldsymbol{\sigma}^{(y)g}}\right) - \boldsymbol{h}^{(L)T}\boldsymbol{W}^{(L+1)g}\left(\frac{\boldsymbol{y}^g}{\boldsymbol{\sigma}^{(y)g} \circ \boldsymbol{\sigma}^{(y)g}}\right) - \boldsymbol{d}^{cT}\boldsymbol{y}^c - \boldsymbol{h}^{(L)T}\boldsymbol{W}^{(L+1)c}\boldsymbol{y}^c
\end{aligned}
\tag{5.3}
$$

ここで，$\boldsymbol{x}^c \in \{0,\ 1\}^{X^c}$，$\boldsymbol{x}^g \in \mathbb{R}^{X^g}$ はそれぞれ可視変数 $\boldsymbol{x}$ のうち，カテゴリカル分布に従うユニット，正規分布に従うユニットを表し，$\boldsymbol{y}^c \in \{0,\ 1\}^{Y^c}$，$\boldsymbol{y}^g \in \mathbb{R}^{Y^g}$ もそれぞれ可視変数 $\boldsymbol{y}$ のうち同様のユニットを表す（ただし $X^g + X^c = I$，$Y^g + Y^c = K$，$\boldsymbol{x} = [\boldsymbol{x}^{gT}\ \boldsymbol{x}^{cT}]^T$，$\boldsymbol{y} = [\boldsymbol{y}^{gT}\ \boldsymbol{y}^{cT}]^T$）．また，$\boldsymbol{W}^{(1)c} \in \mathbb{R}^{X^c \times J_1}$，$\boldsymbol{W}^{(L+1)c} \in \mathbb{R}^{J_L \times Y^c}$，$\boldsymbol{b}^c \in \mathbb{R}^{X^c}$ および $\boldsymbol{d}^c \in \mathbb{R}^{Y^c}$ はそれぞれ可視変数のうちカテゴリカル分布に従うユニットに対応するパラメータを表す．同様に，$\boldsymbol{W}^{(1)g} \in \mathbb{R}^{X^g \times J_1}$，$\boldsymbol{W}^{(L+1)g} \in \mathbb{R}^{J_L \times Y^g}$，$\boldsymbol{b}^g \in \mathbb{R}^{X^g}$ および $\boldsymbol{d}^g \in \mathbb{R}^{Y^g}$ はそれぞれ可視変数のうち正規分布に従うユニットに対応するパラメータを表す．そして，$\boldsymbol{\sigma}^{(x)g} \in \mathbb{R}^{X^g}$，$\boldsymbol{\sigma}^{(y)g} \in \mathbb{R}^{Y^g}$ はそれぞれ可視変数 $\boldsymbol{x}^g$，$\boldsymbol{y}^g$ の偏差を表し，いずれも推定すべきパラメータである．式中の除算は要素ごとの除算を表す．

GCDRM のエネルギー関数の定義より，可視層の条件付き確率分布はそれぞれ

$$
p(x_i^c = 1|\boldsymbol{h}^{(1)}) = \frac{\exp(b_i^c + \boldsymbol{W}_{i:}^{(1)c}\boldsymbol{h}^{(1)})}{\sum_{i'} \exp(b_{i'}^c + \boldsymbol{W}_{i':}^{(1)c}\boldsymbol{h}^{(1)})} \tag{5.4}
$$

$$
p(x_i^g = x|\boldsymbol{h}^{(1)}) = \mathcal{N}\left(x|b_i^g + \boldsymbol{W}_{i:}^{(1)g}\boldsymbol{h}^{(1)}, \sigma_i^{(x)g2}\right) \tag{5.5}
$$

$$
p(y_k^c = 1|\boldsymbol{h}^{(L)}) = \frac{\exp(d_k^c + \boldsymbol{W}_{:k}^{(L+1)cT}\boldsymbol{h}^{(L)})}{\sum_{k'} \exp(d_{d'}^c + \boldsymbol{W}_{:k'}^{(L+1)cT}\boldsymbol{h}^{(L)})} \tag{5.6}
$$

$$
p(y_k^g = y|\boldsymbol{h}^{(L)}) = \mathcal{N}\left(y|d_k^g + \boldsymbol{W}_{:k}^{(L+1)gT}\boldsymbol{h}^{(L)}, \sigma_k^{(y)g2}\right) \tag{5.7}
$$

となる．また，1 層目，L 層目の隠れ層について，条件付き確率分布はそれぞれ

$$p(h_j^{(1)}=1|\boldsymbol{x},\boldsymbol{h}^{(2)}) = \sigma\left(c_j^{(1)}+\boldsymbol{W}_{:j}^{(1)T}\left(\frac{\boldsymbol{x}}{\boldsymbol{\sigma}^{(x)}\circ\boldsymbol{\sigma}^{(x)}}\right)+\boldsymbol{W}_{j:}^{(2)}\boldsymbol{h}^{(2)}\right) \tag{5.8}$$

$$p(h_j^{(L)}=1|\boldsymbol{y},\boldsymbol{h}^{(L-1)}) = \sigma\left(c_j^{(L)}+\boldsymbol{W}_{:j}^{(L)T}\boldsymbol{h}^{(L-1)}+\boldsymbol{W}_{j:}^{(L+1)}\left(\frac{\boldsymbol{y}}{\boldsymbol{\sigma}^{(y)}\circ\boldsymbol{\sigma}^{(y)}}\right)\right) \tag{5.9}$$

となる．ただし，簡単のため結合重み $\boldsymbol{W}^{(1)}$, $\boldsymbol{W}^{(L+1)}$ についてそれぞれ $\boldsymbol{W}^{(1)}\equiv[\boldsymbol{W}^{(1)gT}\ \boldsymbol{W}^{(1)cT}]^T$, $\boldsymbol{W}^{(L+1)}\equiv[\boldsymbol{W}^{(L+1)gT}\ \boldsymbol{W}^{(L+1)cT}]^T$ としている．また，偏差 $\boldsymbol{\sigma}^{(x)},\boldsymbol{\sigma}^{(y)}$ についても簡単のため，それぞれ

$$\sigma_i^{(x)} = \begin{cases}\sigma_i^{(x)g} & (1\le i\le X^g)\\ 1 & (X^g<i\le I)\end{cases} \tag{5.10}$$

$$\sigma_k^{(y)} = \begin{cases}\sigma_k^{(y)g} & (1\le k\le Y^g)\\ 1 & (Y^g<k\le K)\end{cases} \tag{5.11}$$

としている．$2,\ldots,L-1$ 層目の隠れ層の条件付き確率分布は式 (4.28) と同様である．式 (5.5) から式 (5.9) の条件付き確率分布の導出は付録 A を参照のこと．

可視層のバイアス項について $\boldsymbol{b}\equiv[\boldsymbol{b}^{gT}\ \boldsymbol{b}^{cT}]^T$, $\boldsymbol{d}\equiv[\boldsymbol{d}^{gT}\ \boldsymbol{d}^{cT}]^T$ とすれば，GCDRM のパラメータ $\boldsymbol{\theta}=\{\boldsymbol{W},\boldsymbol{b},\boldsymbol{c},\boldsymbol{d},\boldsymbol{\sigma}^{(x)g},\boldsymbol{\sigma}^{(y)g}\}$ は，従来の DRM と同様に対数尤度 $\mathcal{L}$ が最大となるように推定される．それぞれのパラメータに関する勾配は

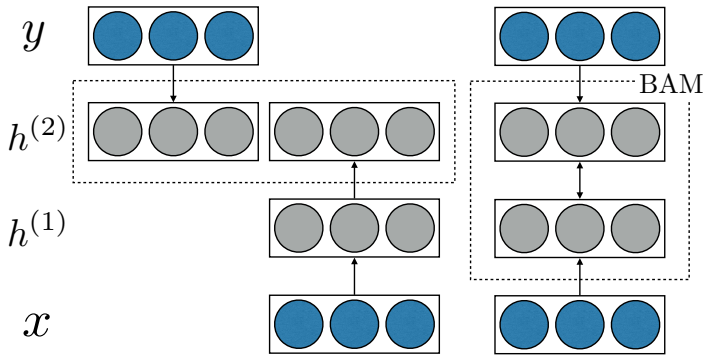
$$\frac{\partial\mathcal{L}}{\partial b_i} = \left\langle\frac{1}{\sigma_i^{(x)2}}x_i\right\rangle_{data}-\left\langle\frac{1}{\sigma_i^{(x)2}}x_i\right\rangle_{model} \tag{5.12}$$

$$\frac{\partial\mathcal{L}}{\partial c_j^{(l)}} = \langle h_j^{(l)}\rangle_{data}-\langle h_j^{(l)}\rangle_{model} \tag{5.13}$$

$$\frac{\partial\mathcal{L}}{\partial d_k} = \left\langle\frac{1}{\sigma_k^{(y)2}}y_k\right\rangle_{data}-\left\langle\frac{1}{\sigma_k^{(y)2}}y_k\right\rangle_{model} \tag{5.14}$$

$$\frac{\partial\mathcal{L}}{\partial W_{ij}^{(l)}} = \begin{cases}\left\langle\frac{1}{\sigma_i^{(x)2}}x_ih_j^{(1)}\right\rangle_{data}-\left\langle\frac{1}{\sigma_i^{(x)2}}x_ih_j^{(1)}\right\rangle_{model} & (l=1)\\ \langle h_i^{(l-1)}h_j^{(l)}\rangle_{data}-\langle h_i^{(l-1)}h_j^{(l)}\rangle_{model} & (l=2,\ldots,L)\\ \left\langle\frac{1}{\sigma_j^{(y)2}}h_i^{(L)}y_j\right\rangle_{data}-\left\langle\frac{1}{\sigma_j^{(y)2}}h_i^{(L)}y_j\right\rangle_{model} & (l=L+1)\end{cases} \tag{5.15}$$

となる．ここで，可視変数に関するデータの期待値 $\langle\boldsymbol{x}\rangle_{data}$, $\langle\boldsymbol{y}\rangle_{data}$ は従来の DRM 同様，観測データの平均を計算することで得られる．また，隠れ変数に関するデータの期待値 $\langle\boldsymbol{h}^{(l)}\rangle_{data}$ は，観測データを与えて式 (4.28) より隠れ層の値を T 回更新することで得られる．ただし，GCDRM の学習では，可視層と隣接する隠れ層の値 $\boldsymbol{h}^{(1)}$, $\boldsymbol{h}^{(L)}$ は，それぞれ式 (5.8)，式 (5.9) によって計算される．一方，モデルの



(a) RBM　　(b) BAM

図 5–1: Pre-training methods of using (a) only RBMs, and (b) RBMs and BAM.

期待値 $\langle\cdot\rangle_{model}$ の計算は組合せ爆発の問題が生じる．そこで，GCDRM の学習においても，平均場近似を用いて近似する．まず，可視変数に関するモデルの期待値は $\langle \boldsymbol{h}^{(l)}\rangle_{data}$ を用いて計算される．その際，$\boldsymbol{x}^c$, $\boldsymbol{x}^g$ の値はそれぞれ式 (5.4)，式 (5.5) より，$\boldsymbol{y}^c$, $\boldsymbol{y}^g$ の値はそれぞれ式 (5.6)，式 (5.7) より計算される．そして，得られた $\langle \boldsymbol{x}\rangle_{model}$, $\langle \boldsymbol{y}\rangle_{model}$ を用いて，$\langle \boldsymbol{h}^{(l)}\rangle_{model}$ を得る．ここでも，可視層と隣接する隠れ層の値は式 (5.8)，式 (5.9) より計算する．

GCDRM のパラメータは式 (5.12) から式 (5.15) を用いて繰返し更新される．また，分散パラメータについては常に非負値となるように，IGBRBM と同様，その対数 $z_i^{(x)} = \log \sigma_i^{(x)2}$, $z_k^{(y)} = \log \sigma_k^{(y)2}$ を更新することで推定する．したがって，それぞれ

$$\frac{\partial \mathcal{L}}{\partial z_i^{(x)}} = e^{-z_i^{(x)}} \left(\left\langle \frac{1}{2}(x_i - b_i)^2 - x_i \boldsymbol{W}_{i:}^{(1)} \boldsymbol{h}^{(1)} \right\rangle_{data} - \left\langle \frac{1}{2}(x_i - b_i)^2 - x_i \boldsymbol{W}_{i:}^{(1)} \boldsymbol{h}^{(1)} \right\rangle_{model} \right) \tag{5.16}$$

$$\frac{\partial \mathcal{L}}{\partial z_k^{(y)}} = e^{-z_k^{(y)}} \left(\left\langle \frac{1}{2}(y_k - d_k)^2 - y_k \boldsymbol{W}_{:k}^{(L+1)T} \boldsymbol{h}^{(L)} \right\rangle_{data} - \left\langle \frac{1}{2}(y_k - d_k)^2 - y_k \boldsymbol{W}_{:k}^{(L+1)T} \boldsymbol{h}^{(L)} \right\rangle_{model} \right) \tag{5.17}$$

と計算される．

5.1.2　GCDRM の事前学習

4.4 節で述べたとおり，従来の DRM は，可視層側から隠れ層側へ，すなわち外側から内側へと RBM を構築し，層ごとに逐次貪欲的に事前学習を行う．2 つの隠れ層間のパラメータを事前学習する際は，1 つ前のステップで得られた RBM からサンプルした隠れ変数を擬似的な可視変数とみなす．このとき，$\boldsymbol{x}$ 側からの RBM と $\boldsymbol{y}$ 側からの RBM が，それぞれ異なる隠れ変数を表すことになる（図 5–1 （a））．それを避けるため，GCDRM の事前学習では，RBM だけでなく，4.2 節で説明した BAM を用いて事前学習を行なう（図 5–1 （b））．

5.1.3　GCDRM を用いた音声認識・音声合成

GCDRM を音声認識および音声合成へ応用する場合，言語特徴量 $\boldsymbol{w}$ を 1 つ目の可視変数 $\boldsymbol{x}$ に，音響特徴量 $\boldsymbol{a}$ を 2 つ目の可視変数 $\boldsymbol{y}$ に割当てる．言語特徴量は one-hot 表現により音素を表すバイナリ値と，コンテキストラベルを表す実数値から構成される．一方，音響特徴量は実数値のみから構成されるため，GCDRM のエネルギー関数および式 (5.6) 中の可視変数 $\boldsymbol{y}^c$ に関する項を無視することで，音声認識および音声合成を行う．

GCDRM の学習を行ったのちに，得られたパラメータを初期値として与えた DNN を構築し，BP 法を用いてパラメータの fine-tuning を行う．その際，音声認識器を構築する場合には，入力を音響特徴量，出力を言語特徴量とする．構築する DNN のバイアスには，学習済みの GCDRM のバイアス $\boldsymbol{d}$, $\boldsymbol{c}^{(L)}$, …, $\boldsymbol{c}^{(1)}$, $\boldsymbol{b}$ を入力層から出力層へと順に割当てる．また，DNN の結合重みには，GCDRM の結合重みをそれぞれ転置した $\boldsymbol{W}^{(L+1)T}$, $\boldsymbol{W}^{(L)T}$, …, $\boldsymbol{W}^{(1)T}$ を入力層に近い結合から順に割当てる．

一方，音声合成器を構築する場合には，入力を言語特徴量，出力を音響特徴量とする．構築する DNN のパラメータの初期値として，バイアスには，学習済みの GCDRM のバイアス $\boldsymbol{b}$, $\boldsymbol{c}^{(1)}$, …, $\boldsymbol{c}^{(L)}$, $\boldsymbol{d}$ を入力層から出力層へと順に割当てる．また，DNN の結合重みには，GCDRM の結合重み $\boldsymbol{W}^{(1)}$, $\boldsymbol{W}^{(2)}$, …, $\boldsymbol{W}^{(L+1)}$ を入力層に近い結合から順に割当てる．GCDRM は言語特徴量と音響特徴量の結合分布をモデル化しているため，認識器あるいは合成器のどちらを構築する場合でも，DNN の初期値として同じパラメータを与えることができる．

5.2　実験

5.2.1　実験条件

本実験では，HTS ワーキンググループ [42] が公開している，日本語話者による音声データセットである NIT ATR503 M001 を用いて，提案法を用いた音声認識および音声合成の精度を評価した．ただし，本実験では音声認識の基本単位として音素を採用した．このデータセットは，サンプリング周波数が 48kHz であり，単一の男性話者によって，音素バランスがとれた文章を読上げた音声と，読上げた文章を形態素解析したコンテキストラベルデータが含まれている．データセットはセット A からセット J までの 10 セットで構成され，セット A からセット I まではそれぞれ 50 文，セット J は 53 文を含んでいる．このデータセット内で使用されている音素の種類は 42 種類であり，また，コンテキストラベルの種類は単語の活用形を表す ID や，フレーズのアクセントタイプを表す ID などからなる 45 種類である．本実験では，当該音素のフレーム長および当該音素における当該フレームの相対位置の 2 つをコンテキストラベルに加え，合計 47 種類のコンテキストラベルを使用した．本実験で使用した音素ラベル，コンテキストラベルの一覧はそれぞれ付録 B，C を参照のこと．

実験に使用する言語特徴量として，先行，当該，後続の 3 音素を表す one-hot ベクトルおよびコンテキストラベルを用いた．また，音響特徴量としてメルケプストラムを用いた．メルケプストラムの次元数は 35 とし，さらにその動的特徴量（一次微分および二次微分）を用いた．したがって，言語特徴量の次元数は 173（$= 42 \times 3 + 47$），音響特徴量の次元数は 135（$= 35 \times 3$）となる．学習データは，言語特徴量のうちコンテキストラベル，音響特徴量について，それぞれ次元ごとに平均 0，分散 1 に標準化を行った．

5.2.2　音声合成タスク

図 5–2 は自然音声，DNN の出力および提案法の出力より得られた音響特徴量から計算された 8 次のメルケプストラムの軌跡をプロットしたものである．図中の“ GCDRM ”が提案法を示す．図から，提案法が DNN よりも，より自然音声に近い音響特徴量を生成できていることが確認できる．

提案法の最適なパラメータ数を決定するため，隠れ層の層数および各隠れ層のユニット数を変化させ，音声合成の精度を比較した．実験はデータセット中のセット

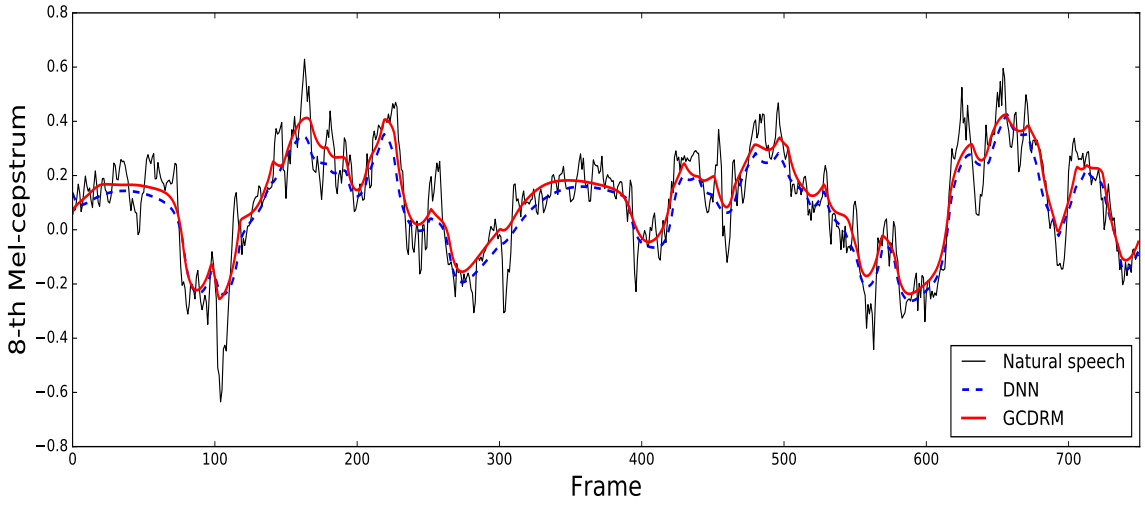


図 5–2: Trajectories of 8-th Mel-cepstral coefficients of natural speech and those generated by the DNN and the proposed systems.

A および B に含まれる合計 100 文を用いて学習し，セット J に含まれる 53 文を用いてテストを行なった．言語特徴量を入力として得られた GCDRM の出力を平均，全学習データから計算した分散を共分散行列として，Maximum Likelihood Parameter Generation（MLPG） [43] を用いてメルケプストラムを出力した．音声合成の精度を評価する指標として，次の式で表されるメルケプストラム歪み（Mel-cepstral distortion; MCD） [19] を使用した．

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(mc_d^t - mc_d^e)^2} \tag{5.18}$$

ただし，mc_{data}^t, mc_{data}^e はそれぞれ自然音声，モデルが生成した音声のあるフレームにおけるメルケプストラムの d 次元目の値である．また，本実験において音響特徴量として使用したメルケプストラムの次数は 35 であるため，式 (5.18) において $D = 35$ とした．MCD は，自然音声とモデルが生成した音声の，メルケプストラム間のユークリッド距離を表す．したがって，MCD が低いほどターゲットとなる自然音声に近い音声を生成できていることになり，精度が良いことを示す．テストの際は，セット J に含まれる 53 文に対して，正解となる言語特徴量を入力して提案法から得られた出力と，正解となる音響特徴量から計算される MCD の平均を比較した．

実験の結果を図 5–3 に示す．図において，例えば (3×200) はユニット数が 200

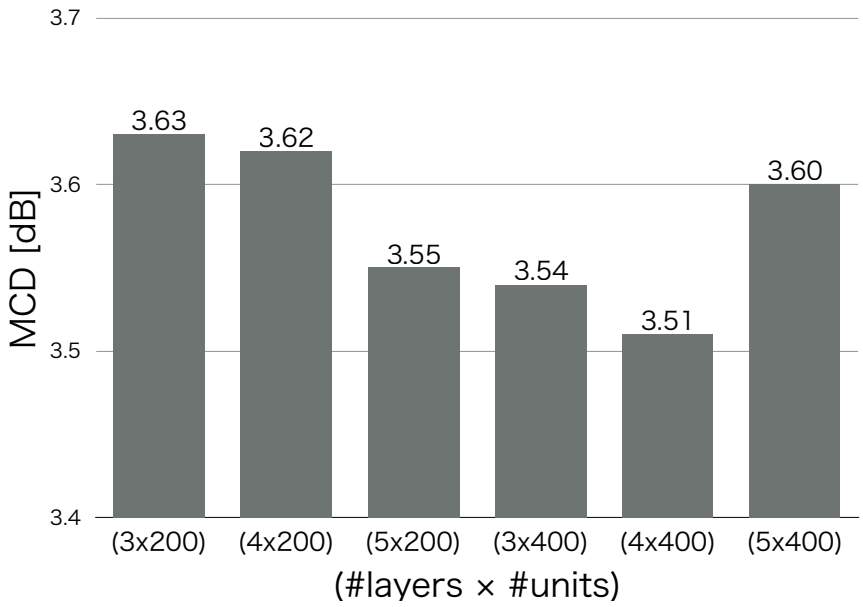


図 5–3: Performance of our method when changing the number of hidden layers and hidden units at each hidden layer (MCD [dB]).

の隠れ層が 3 層で構成されるモデルを示す．図 5–3 から，(4 × 400) のモデルが最も MCD が低く，良い結果を示すことがわかる．また，ユニット数が 200 のときは，層数を増やすと精度が上がることがわかる．しかし，ユニット数が 400 のときは，層数を 5 まで増やすと精度が下がる．これは，学習時に推定すべきパラメータの数が増加することで，モデルがうまく学習できないことが原因であると考えられる．

客観評価

次に，提案法による音声合成の性能を客観的に評価するため，MCD を用いて DNN，DBN による結果と比較した．学習データに用いる発話文の数を 50 文，100 文，150 文，200 文，450 文と変化させ，それぞれの学習データ数についてテストし，比較を行った．テストには学習データに含まれない 53 文を使用した．それぞれの手法について，(4 × 400) のモデルを用いた．DBN および GCDRM は，学習したパラメータ（結合重みおよびバイアス）を DNN のパラメータの初期値として与えて fine-tuning を行なった．また，DNN のパラメータはランダムな初期値を与えて学習を行なった．

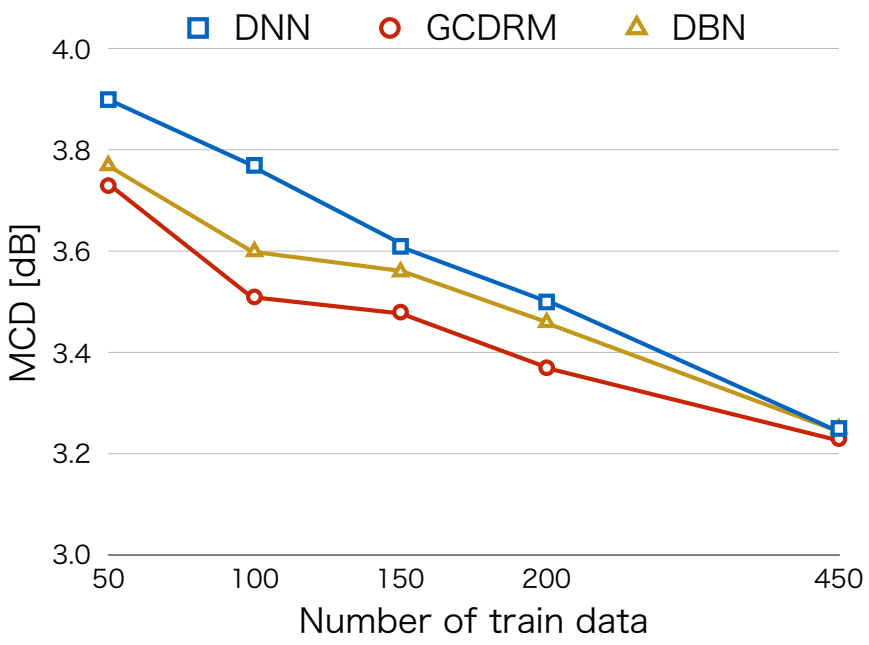


図 5–4: Comparison of MCD [dB] between the generated speech and the target speech obtained by each method.

実験の結果を図 5–4 に示す．図 5–4 から，いずれの学習データ数においても，提案法が合成精度を改善していることがわかる．学習データ数が 450 文のときは各手法についてほとんど結果の差がない．これは，DNN の学習において，学習データを十分に与えればパラメータをうまく最適化することができることが原因であると考えられる．したがって，提案法は学習データ数が少ないときにとくに有用であるといえる．これは，提案法言語特徴量と音響特徴量の結合分布をモデル化していることが理由であると考えられる．DNN および DBN は言語特徴量（入力）から音響特徴量（出力）への単一方向の関係性のみを表現しているのに対して，提案法は言語特徴量から音響特徴量だけでなく，音響特徴量から言語特徴量の，双方向の関係性を表現している．したがって，提案法の学習では，言語特徴量から生成される音響特徴量が自然であるかだけでなく，生成された音響特徴量が言語特徴量として正しく認識されるかも考慮してパラメータを最適化していると考えられる．

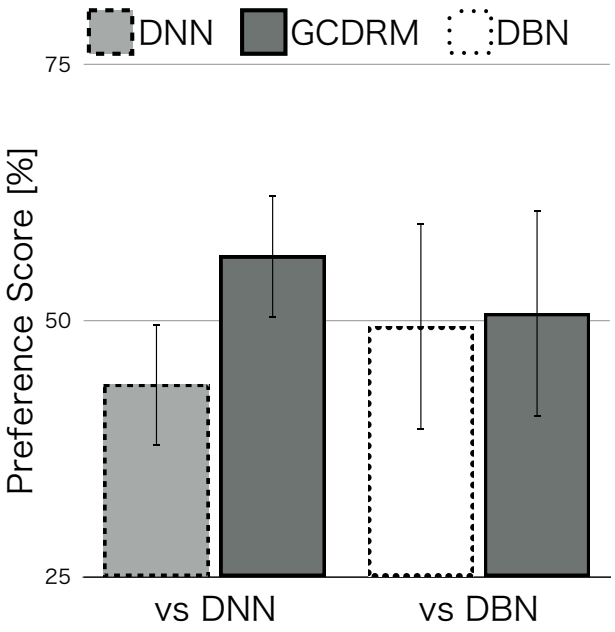


図 5–5: Subjective preference scores [%] of speech samples obtained by each method.

主観評価

次に，提案法による音声合成の性能を主観的に評価するために，200 文で学習した DNN，DBN および提案法から得られた合成音声の自然性を，XAB テストにより評価した．テストに使用した音声は，学習データに含まれない 53 文の中からランダムに選んだ 20 文であり，比較したペアは DNN と提案法，DBN と提案法の 2 ペアである．いずれの手法も (4×400) のモデルを用いた．また，音声の基本周波数および状態継続長は自然音声から抽出したものを使用し，メルケプストラムのみ，各手法から出力されたものを使用して合成音声を生成した．被験者は 20 代の学生 8 名であり，被験者には各サンプル音声を聞いたのち，どちらがより自然音声に近いかを選択してもらった．

XAB テストによる合成音声の自然性に関する評価結果を図 5–5 に示す．図中の誤差範囲は 95%信頼区間を示す．図 5–5 の結果から，DNN と提案法では，提案法が有効であることがわかる．一方，DBN と提案法では，わずかに提案法のほうが良い結果が得られているが，両手法の自然性に 5%の有意水準で有意差は認められなかった．しかし，提案法は DBN と違い，音声合成器だけでなく音声認識器も同時に事前学習することができる点において有用であるといえる．音声認識実験の結果は次の節で述べる．

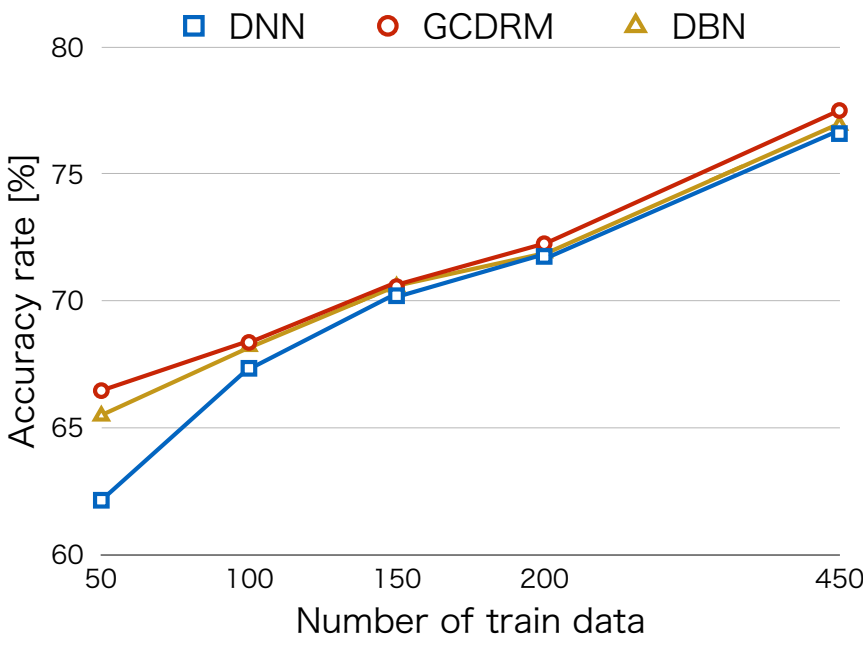


図 5–6: Accuracy rate [%] of the current phoneme obtained by each method.

5.2.3　音声認識タスク

最後に，提案法による音声認識の性能を客観的に評価するため，音素認識実験を行なった．実験は，音響特徴量を入力として推定された言語特徴量の中から，当該音素の正解率を評価した．評価の方法として，学習データ数を 50 文，100 文，150 文，200 文，450 文と変化させ，それぞれの学習データ数についてテストした結果を，提案法，DNN および DBN について比較した．テストには学習データに含まれない 53 文を使用し，53 文の平均正解率で評価を行なった．また，それぞれの手法について (4 × 400) のモデルを用いた．提案法では，5.2.2 節の音声合成実験で学習した GCDRM のパラメータと同じ値を使用し，5.1.3 節で説明した方法で DNN の初期値として与えたのち，fine-tuning を行った．また，DBN および DNN は 5.2.2 節の実験と同様の条件で学習した．

実験の結果を図 5–6 に示す．図 5–6 から，いずれの学習データ数においても，認識精度が向上していることがわかる．しかし，学習データ数が 150 文，200 文以降は結果に大きな差が見られない．これは 5.2.2 節と同様の理由によると考えられる．

5.3　まとめ

本章では，DRM を音声認識および音声合成に応用するため，正規分布およびカテゴリカル分布へと拡張した GCDRM を定義した．その後，GCDRM を用いた音声認識および音声合成手法について述べた．そして，実験によって GCDRM を用いて音声認識および音声合成を行うことが可能であることを示した．また，客観評価実験において，提案法が従来法である DNN，DBN よりも認識・合成精度を改善した．とくに，学習データ数が少ないときに精度を改善した．音声合成タスクにおける主観評価実験では，提案法が DNN よりも有効であることがわかった．

第6章

DRMの声質変換への応用

本章では，ソース話者からターゲット話者・ターゲット話者からソース話者へ，双方向な声質変換を実現する声質変換器を構築するために，DRM を用いてソース話者とターゲット話者の結合確率を表現する手法を提案する．声質変換に用いる特徴量である音響特徴量は実数値のみから構成される．したがって，4.4 節で説明した従来の DRM のように Bernoulli 分布ではなく，音響特徴量を表す正規分布を扱う必要がある．そこで，従来の DRM を正規分布へと拡張した Gaussian-Gaussian DRM（GGDRM）を定義する．また，GGDRM を用いた声質変換手法について述べる．そして，GGDRM を用いて声質変換を行うことが可能であることを実験によって示す．

6.1 Gaussian-Gaussian DRM

6.1.1 GGDRM の定義

正規分布のみを扱う GGDRM は，第 5 章で述べた GCDRM の特殊な場合として定義される．すなわち，GGDRM のエネルギー関数は式 (5.3) の GCDRM のエネルギー関数から，カテゴリカル分布に対応する変数やパラメータをすべて無視して

$$
\begin{aligned}
E(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \boldsymbol{\theta}) \quad = \quad & \frac{1}{2}\left(\frac{\boldsymbol{x}-\boldsymbol{b}}{\boldsymbol{\sigma}^{(x)}}\right)^T\left(\frac{\boldsymbol{x}-\boldsymbol{b}}{\boldsymbol{\sigma}^{(x)}}\right) - \left(\frac{\boldsymbol{x}}{\boldsymbol{\sigma}^{(x)} \circ \boldsymbol{\sigma}^{(x)}}\right)^T \boldsymbol{W}^{(1)}\boldsymbol{h}^{(1)} \\
& - \sum_{l=1}^{L} \boldsymbol{c}^{(l)T}\boldsymbol{h}^{(l)} - \sum_{l=2}^{L} \boldsymbol{h}^{(l-1)T}\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l)} \\
& + \frac{1}{2}\left(\frac{\boldsymbol{y}-\boldsymbol{d}}{\boldsymbol{\sigma}^{(y)}}\right)^T\left(\frac{\boldsymbol{y}-\boldsymbol{d}}{\boldsymbol{\sigma}^{(y)}}\right) - \boldsymbol{h}^{(L)T}\boldsymbol{W}^{(L+1)}\left(\frac{\boldsymbol{y}}{\boldsymbol{\sigma}^{(y)} \circ \boldsymbol{\sigma}^{(y)}}\right) \qquad (6.1)
\end{aligned}
$$

と定義される．

そして，GGDRM の可視層の条件付き確率分布は，$\boldsymbol{x}^g = \boldsymbol{x}$, $\boldsymbol{y}^g = \boldsymbol{y}$, $\boldsymbol{W}^{(1)g} = \boldsymbol{W}^{(1)}$, $\boldsymbol{W}^{(L+1)g} = \boldsymbol{W}^{(L+1)}$, $\boldsymbol{b}^g = \boldsymbol{b}$, $\boldsymbol{d}^g = \boldsymbol{d}$ とみなして，式 (5.4) および式 (5.7) より与えられる．また，隠れ層の条件付き確率分布およびそれぞれのパラメータに関する勾配は GCDRM と同様に計算される．

6.1.2　GGDRM を用いた声質変換

GGDRM を声質変換へ応用する場合，話者 X の音響特徴量 $\boldsymbol{a}^X$ を 1 つ目の可視変数 $\boldsymbol{x}$，別の話者 Y の音響特徴量 $\boldsymbol{a}^Y$ を 2 つ目の可視変数 $\boldsymbol{y}$ とみなす．

GGDRM の学習を行ったのちに，得られたパラメータを初期値として与えた DNN を構築し，BP 法を用いてパラメータの fine-tuning を行う．その際，話者 X から話者 Y へ声質変換を行う変換器を構築する場合は，入力を音響特徴量 $\boldsymbol{a}^X$，出力を音響特徴量 $\boldsymbol{a}^Y$ とする．一方，話者 Y から話者 X へ声質変換を行う変換器を構築する場合は，入出力に割当てる音響特徴量を入替える．また，構築する DNN のパラメータは，5.1.3 節で述べた GCDRM の場合と同様に割当てる．

6.2　実験

6.2.1　実験条件

本実験では，Voice Conversion Challenge 2018 [44] で公開された，英語話者による音声データセットを用いた．データセットは，サンプリング周波数が 22050Hz であり，女性 4 名，男性 4 名の計 8 名がそれぞれ同じ文章を読み上げた 81 文から構成される（各文は 3 秒程度）．学習データとして，全ての話者に共通する 50 文を使用し，残りをテストデータをして用いた．学習の際は，音声を 24kHz にアップサンプリングしたものを使用した．また，DP マッチングによって，話者ペアごとに発話の長さを揃えた音声を使用した．学習する話者ペアは，同一性別が男女それぞれ 8 ペア，異性別が，女性から男性，男性から女性についてそれぞれ 8 ペアである（計 32 ペア）．

実験に使用する音響特徴量としてメルケプスラムを用いた．メルケプストラムの次数は 40 とし，さらにその動的特徴量（一次微分および二次微分）を考慮した．

提案法と比較する従来法として，GMM および DNN による実験を行なった．提

案法および DNN はユニット数 600 からなる隠れ層が 3 層のモデルを用いた．提案法では，GGDRM による学習を行ったのちに，得られたパラメータ（結合重みおよびバイアス）を DNN のパラメータ初期値として与えて fine-tuning を行なった．DNN のパラメータの初期値はランダムな値を与えて学習を行なった．ある話者ペア（話者 X，話者 Y）について学習を行う際，話者 X から話者 Y への変換と，話者 Y から話者 X への変換について，提案法では DNN に与える初期値として同じ値を使用した．また，GMM の混合数は 64 とした．

テストの際は，各手法から得られた音響特徴量を平均，全学習データから計算した分散を共分散行列として，MLPG を用いてメルケプストラムを出力した．また，声質変換の精度を評価する指標として，式 (5.18) で表される MCD を用いた．ただし，本実験で使用したメルケプストラムの次数が 40 であるため，$D = 40$ とした．MCD を計算する際の前処理として，各手法により得られた変換音声と，ターゲットとなる話者の自然音声で DP マッチングを行い，発話の長さを揃えた．学習データは，音響特徴量をそれぞれ次元ごとに平均 0，分散 1 に標準化を行った．

6.2.2　実験結果

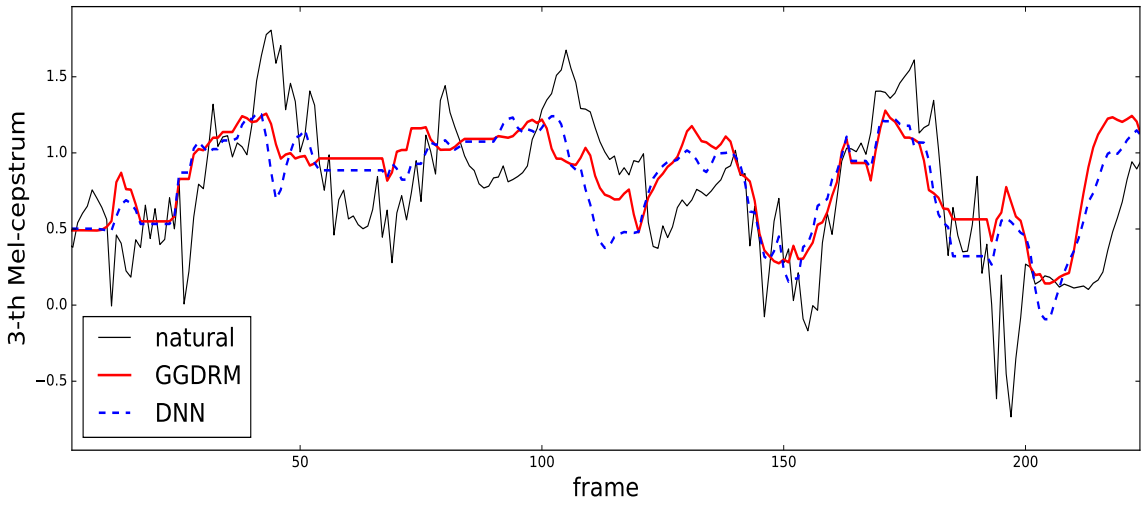


図 6–1: Trajectories of 3-th Mel-cepstral coefficients of natural speech and those generated by the DNN and the proposed systems.

最初に，図 6–1 に，男性から女性への声質変換実験において DNN および提案法より出力された音響特徴量から得られた，3 次のメルケプストラムの軌跡を示す．

表 6–1: Comparison of MCD [dB] obtained by each method. For example, “ f2m ” indicates female-to-male conversion.

	MCD [dB]			
	f2f	m2m	f2m	m2f
GMM	6.21	6.16	6.41	6.37
DNN	5.53	5.48	5.59	5.62
GGDRM	**5.43**	**5.36**	**5.48**	**5.41**

図中の“ GGDRM ”が，提案法を示す．提案法により得られた軌跡のほうが，DNN による軌跡よりも，ターゲット音声に近いことがわかる．

客観評価

次に，各手法による声質変換実験の結果を表 6–1 に示す．表 6–1 より，同性間の実験結果（“ f2f ”および“ m2m ”）から，提案法が，従来法よりも声質変換の性能を向上させていることがわかる．また，異性間の声質変換実験の結果（“ f2m ”および“ m2f ”）においても，提案法が従来法よりも良い性能を示すことがわかる．提案法では，GGDRM を用いて，ソース話者の音声からターゲット話者の音声への単一方向の関係性だけでなく，変換した音声からソース話者の音声へ再変換するような逆向きの関係性を考慮して学習することで，話者間の深い関係性を適切に表現することができたと考えられる．

主観評価

提案法の性能を主観的に評価するため，11 人の被験者を対象に聴取実験を行った．本実験では，31 文のテストデータからランダムに選んだ 10 文を使用して，GMM と提案法，DNN と提案法を比較した．被験者には，類似性（どちらのモデルの音声が，ターゲット話者の自然音声に似ているか）および品質（どちらのモデルの音声が，よりクオリティが高いか）の 2 つの基準で評価してもらった．テストに使用する音声波形は，モデルから出力された音響特徴量と，次の式で線形に変換

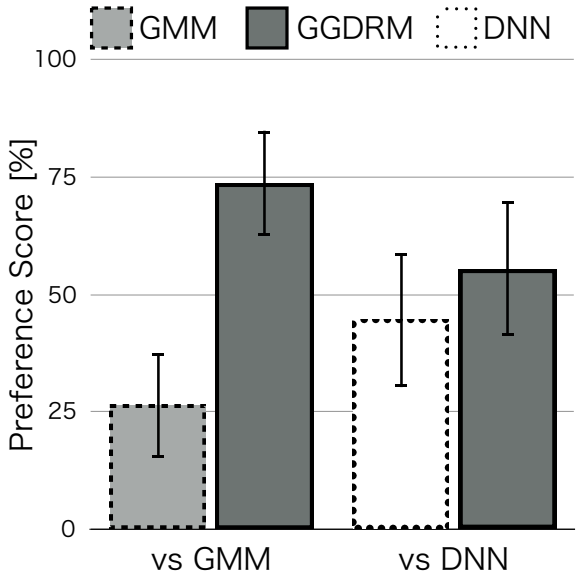


図 6–2: Subjective preference scores [%] for similarity of speech samples obtained by each method.

された F0 から合成した．

$$\hat{y}_t = \frac{\rho^{(y)}}{\rho^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)} \tag{6.2}$$

ここで，x_t, $\hat{y}_t$ はそれぞれソース話者の音声，変換された音声のログスケールにおける F0 の値である．そして，$\rho^{(x)}$, $\mu^{(x)}$ はそれぞれ学習データ中に含まれるソース話者の音声から計算される F0 の標準偏差および平均を表し，$\rho^{(y)}$, $\mu^{(y)}$ はそれぞれターゲット話者について同様の値を表す．また，t はフレームを表すインデックスである．

類似性に関する主観評価実験の結果を図 6–2 に示す．図中の誤差範囲は 95%信頼区間を表す．図より，提案法が GMM よりも性能を改善していることがわかる．一方，DNN との比較では有意差が認められなかった．しかし，ソース話者からターゲット話者，ターゲット話者からソース話者への変換器を構築する際に，DNN の初期値として同じ値を与えることができる提案法が，ソース話者からターゲット話者への変換器のみを構築する DNN に匹敵する性能を示すことがわかる．

また，図 6–3 に示す音声の品質に関する主観評価実験も，類似性に関する実験と同様結果であることがわかる．したがって，提案法は声質変換に有効であることがわかる．

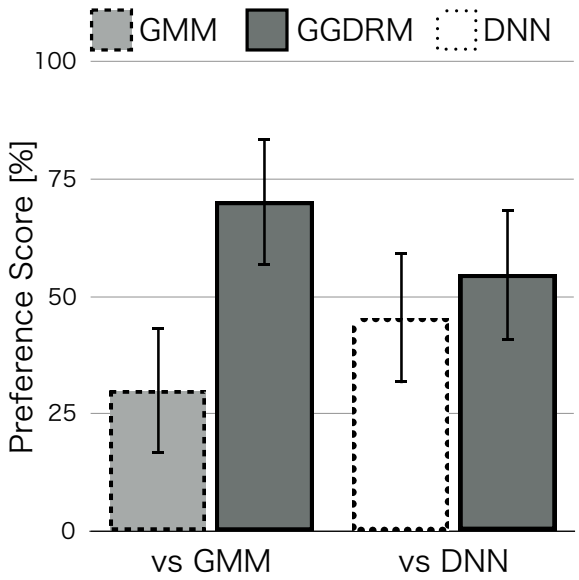


図 6–3: Subjective preference scores [%] of for quality speech samples obtained by each method.

6.3　まとめ

本章では，DRM を声質変換に応用するため，GCDRM の特殊な場合である GGDRM について説明した．その後，GGDRM を用いた声質変換手法について述べた．そして，実験によって，GGDRM を用いて声質変換を行うことが可能であることを示した．また，提案法が従来法である GMM，DNN よりも声質変換精度を改善することを示した．

第7章

まとめ

本論文では，深い構造をもつネットワークを用いて 2 つの可視変数間の双方向変換を実現する DRM に着目し，DRM を音声合成処理へ応用する手法を提案した．そして，音声認識，音声合成および声質変換実験によりその有用性を示した．まず，第 5 章では，DRM を音声認識および音声合成へと応用するため，DRM をカテゴリカル分布および正規分布へと拡張した GCDRM を定義した．そして，GCDRM を用いた音声認識および音声合成について説明した．音声認識および音声合成の客観評価実験において，GCDRM を用いた提案法が，テキスト・音声間の双方向の関係性を表現することで，音声認識および音声合成を行うことが可能であることを示した．また，従来法である DNN，DBN を用いた手法よりも認識精度および合成精度を改善することを示した．とくに，学習データ数が少ないときに精度を改善することが確認された．音声合成タスクにおける主観評価実験においては，提案法が DNN よりもより自然な音声を生成できることを示した．また，音声認識器と音声合成器を同時に構築する提案法が，音声合成器のみを構築する DBN と同程度の自然性をもつ音声を生成できることを示した．次に，第 6 章において，DRM を声質変換へと応用するため，GCDRM の特殊な場合である GGDRM について述べた．そして，GGDRM を用いた声質変換について説明した．声質変換実験において，GGDRM を用いた提案法がソース・ターゲット話者間の双方向の関係性を表現することで，声質変換を行うことが可能であることを示した．また，従来法である GMM，DNN を用いた手法よりも声質変換精度を向上することを示した．今後は，特徴量抽出や，声質変換における DP マッチングといった前処理を行うことなく，音声・テキストや，異なる話者の発話を双方向に変換可能な手法について検討したい．

参考文献

[1] B. H. Juang, and L. R. Rabiner,“ Hidden Markov models for speech recognition, ” *Technometrics*, vol. 33, no. 3, pp. 251-272, 1991.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury,“ Deep neural networks for acoustic modeling in speech recognition, ”*IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[3] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn,“ Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, ” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4277-4280, 2012.

[4] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura,“ Speech synthesis based on hidden Markov models, ”*Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234-1252, 2013.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura,“ Simultaneous modeling of spectrum pitch and duration in HMM-based speech synthesis, ” in *Proceedings of the Eurospeech*, pp. 2347-2350, 1999.

[6] A. J. Hunt, and A. W. Black,“ Unit selection in a concatenative speech synthesis system using a large speech database, ”in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 373-376, 1996.

[7] H. Zen, K. Tokuda, and A. W. Black,“ Statistical parametric speech synthesis, ” *Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.

[8] H. Zen, A. Senior, and M. Schuster,“ Statistical parametric speech synthesis using deep neural networks, ”in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7962-7966, 2013.

[9] A. den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu,“ Wavenet: A generative model for raw audio, ”*arXiv:1609.03499*, 2016.

[10] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, “ SampleRNN: An unconditional end-to-end neural audio generation model, ” *arXiv:1612.07837*, 2016.

[11] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, “ Char2wav: End-to-end speech synthesis, ” in *Proceedings of the ICLR2017 workshop (submission)*, 2017.

[12] T. Nakashika, “ Deep relational model: A joint probabilistic model with a hierarchical structure for bidirectional estimation of image and labels, ” *IEICE Transactions on Information and Systems*, vol. E101-D, no. 2, pp. 428-436, 2018.

[13] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “ Voice conversion through vector quantization, ” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 655-658, 1988.

[14] A. Kain, and M.W. Macon, “ Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction, ”in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 813-816, 2001.

[15] H. Valbret, E. Moulines, and J. P. Tubach, “ Voice transformation using PSOLA technique, ” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 145-148, 1992.

[16] L. J. Liu, L. H. Chen, Z. H. Ling, and L. R. Dai, “ Using bidirectional associative memories for joint spectral envelope modeling in voice conversion, ” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7884-7888, 2014.

[17] Y. Stylianou, O. Cappe, and E. Moulines, “ Continuous probabilistic transform for voice conversion, ” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131-142, 1998.

[18] AR. Toth, and A. W. Black, “ Using articulatory position data in voice transformation, ”in *Proceedings of the ISCA Workshop on Speech Synthesis*, pp. 182-187, 2007.

[19] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “ Voice conversion using artificial neural networks, ”in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3893-3896, 2009.

[20] T. Nakashika, T. Takiguchi, and Y. Ariki, “ Voice conversion using RNN pre-trained recurrent temporal restricted Boltzmann machines, ”*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 580-587, 2015.

[21] S. H. Mohammadi, and A. Kain, “ Voice conversion using deep neural networks with speaker-independent pre-training, ”in *Proceedings of the IEEE Spoken Language Technology Workshop*, pp. 19-23, 2014.

[22] 今井聖, 音声信号処理, 森北出版, 東京, 1996.

[23] 古井貞熙, 音響・音声工学, 近代科学社, 東京, 1992.

[24] D. G. Childers, D. P. Skinner, and R. C. Kemerait, “ The cepstrum: A guide to processing, ”*Proceedings of the IEEE*, vol. 65, no. 10, pp. 1428-1443, 1977.

[25] H. Sakoe, and S. Chiba, “ Dynamic programming algorithm optimization for spoken word recognition, ”*IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43-49, 1978.

[26] Y. Bengio, “ Learning deep architectures for AI, ”*Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1-127, 2009.

[27] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai.,“ Speech synthesis from HMMs using dynamic features, ”in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 389-392, 1996.

[28] 岡谷貴之, 深層学習, 講談社, 東京, 2015.

[29] D. H. Ackley, and G. E. Hinton,“ A learning algorithm for Boltzmann machines, ”*Cognitive Science*, vol. 9, no. 1, pp. 147-169, 1985.

[30] G. E. Hinton, S. Osindero, and YW. Teh, “ A fast learning algorithm for deep belief nets, ”*Neural computation*, vol. 18, no. 7, pp. 1527-1554, 2006.

[31] T. Yamashita, M. Tanaka, E. Yoshida, Y. Yamauchi, and H. Fujiyoshi, “ To be Bernoulli or to be Gaussian, for a restricted Boltzmann machine, ”in *Proceedings of the IAPR International Conference on Pattern Recognition (ICPR)*, pp. 1520-1525, 2014.

[32] G. E. Hinton, and R. R. Salakhutdinov,“ Reducing the dimensionality of data with neural networks, ” *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[33] ZH. Ling, L. Deng, and D. Yu, “ Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis, ” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7825-7829, 2013.

[34] KH. Cho, A. Ilin, and T. Raiko, “ Improved learning of Gaussian-Bernoulli restricted Boltzmann machines, ”in *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, pp. 10-17, 2011.

[35] G. E. Hinton,“ Training products of experts by minimizing contrastive divergence, ” *Neural Computation*, vol. 14, no. 8, pp. 1771-1800, 2002.

[36] B. Kosko,“ Bidirectional associative memories, ” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 49-60, 1988.

[37] LH. Chen, T. Raitio, C. Valentini-Botinhao, ZH. Ling, and J. Yamagishi,“ A deep generative architecture for postfiltering in statistical parametric speech synthesis, ” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2003-2014, 2015.

[38] Y. Bengio, P. Lamblim, D. Popovici, and H. Larochelle,“ Greedy layer-wise training of deep nets, ”in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 153-160, MIT Press, 2007.

[39] Y. Tang, “ Deep learning using linear support vector machines, ” in *Proceedings of the ICML Workshop on Challenges in Representation Learning*, 2013.

[40] A. Mohamed, G. E. Dahl, and G. E. Hinton, “ Deep belief networks for phone recognition, ”in *Proceedings of the NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[41] S. Kang, S. Qian, and H. Meng,“ Multi-distribution deep belief network for speech synthesis, ” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8012-8016, 2013.

[42] HTS, http://hts.sp.nitech.ac.jp/

[43] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura,“ Speech parameter generation algorithms for HMM-based speech synthesis, ”in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1315-1318, 2009.

[44] Voice Conversion Challenge 2018, http://www.vc-challenge.org/

謝辞

本研究を行うにあたり，研究の場を与えていただき，なおかつ多くのご指導とご助言をいただいた中鹿亘助教，南泰浩教授ならびに古賀久志准教授に心より深く感謝いたします．日頃から研究に関して活発なご意見，ご助言をいただいた戸田貴久助教に深く感謝いたします．ご多忙の中，多くのご助言，ご協力をくださった柳生智彦客員准教授と鈴木一哉客員准教授に深く感謝いたします．そして，研究室での生活や研究の様々な場面でご助言，ご協力をいただきました南・古賀・戸田・中鹿研究室の学生の皆さま，すでにご卒業された先輩方に心から感謝いたします．

平成 30 年 1 月 29 日

図一覧

表一覧

付録

A GCDRM における条件付き確率分布の導出

GCDRM のエネルギー関数は

$$
\begin{aligned}
E(\boldsymbol{x}^c, \boldsymbol{x}^g, \boldsymbol{y}^c, \boldsymbol{y}^g, \forall \boldsymbol{h}^{(l)}; \boldsymbol{\theta}) \quad = \\
&\frac{1}{2}\left(\frac{\boldsymbol{x}^g - \boldsymbol{b}^g}{\boldsymbol{\sigma}^{(x)g}}\right)^T\left(\frac{\boldsymbol{x}^g - \boldsymbol{b}^g}{\boldsymbol{\sigma}^{(x)g}}\right) - \left(\frac{\boldsymbol{x}^g}{\boldsymbol{\sigma}^{(x)g} \circ \boldsymbol{\sigma}^{(x)g}}\right)^T \boldsymbol{W}^{(1)g}\boldsymbol{h}^{(1)} - \boldsymbol{b}^{cT}\boldsymbol{x}^c - \boldsymbol{x}^{cT}\boldsymbol{W}^{(1)c}\boldsymbol{h}^{(1)} \\
&- \sum_{l=1}^{L} \boldsymbol{c}^{(l)T}\boldsymbol{h}^{(l)} - \sum_{l=2}^{L} \boldsymbol{h}^{(l-1)T}\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l)} \\
&+ \frac{1}{2}\left(\frac{\boldsymbol{y}^g - \boldsymbol{d}^g}{\boldsymbol{\sigma}^{(y)g}}\right)^T\left(\frac{\boldsymbol{y}^g - \boldsymbol{d}^g}{\boldsymbol{\sigma}^{(y)g}}\right) - \boldsymbol{h}^{(L)T}\boldsymbol{W}^{(L+1)g}\left(\frac{\boldsymbol{y}^g}{\boldsymbol{\sigma}^{(y)g} \circ \boldsymbol{\sigma}^{(y)g}}\right) - \boldsymbol{d}^{cT}\boldsymbol{y}^c - \boldsymbol{h}^{(L)T}\boldsymbol{W}^{(L+1)c}\boldsymbol{y}^c
\end{aligned}
$$

と定義されている．GCDRM のエネルギー関数から，正規分布に従う可視変数，カテゴリカル分布に従う可視変数および可視層と隣接する隠れ層の条件付き確率分布を導出する．

正規分布に従う可視変数の条件付き確率分布

まず，$\boldsymbol{x}^g$ の条件付き確率分布を導出する．条件付き確率分布 $p(\boldsymbol{x}^g|\boldsymbol{h}^{(1)})$ は定義より

$$
p(\boldsymbol{x}^g|\boldsymbol{h}^{(1)}) = \frac{p(\boldsymbol{x}^g, \boldsymbol{h}^{(1)})}{p(\boldsymbol{h}^{(1)})} = \frac{p(\boldsymbol{x}^g, \boldsymbol{h}^{(1)})}{\int_{\boldsymbol{x}^g} p(\boldsymbol{x}^g, \boldsymbol{h}^{(1)})}
$$

であるから，分母と分子にそれぞれ GCDRM の結合確率分布から $p(\boldsymbol{x}^g, \boldsymbol{h}^{(1)}) = \exp\{-E(\boldsymbol{x}^g, \boldsymbol{h}^{(1)}; \boldsymbol{\theta})\}/Z$ を代入して

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{h}) \quad &= \quad \frac{\exp\{-E(\boldsymbol{x}, \boldsymbol{h})\}}{\int_{\boldsymbol{x}} \exp\{-E(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta})\} d\boldsymbol{x}} \\
&= \quad \frac{\exp\left(-\sum_i\left(\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_j \frac{W_{ij}x_i h_j}{\sigma_i^2}\right) + \sum_j c_j h_j\right)}{\int_{\boldsymbol{x}} \exp\left(-\sum_i\left(\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_j \frac{W_{ij}x_i h_j}{\sigma_i^2}\right) + \sum_j c_j h_j\right) d\boldsymbol{x}} \\
&= \quad \frac{\prod_i \exp\left(-\left(\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_j \frac{W_{ij}x_i h_j}{\sigma_i^2}\right) + \sum_j c_j h_j\right)}{\prod_i \int_{\boldsymbol{x}} \exp\left(-\left(\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_j \frac{W_{ij}x_i h_j}{\sigma_i^2}\right) + \sum_j c_j h_j\right) d\boldsymbol{x}}
\end{aligned}
$$

を得る．ただし，簡単のため $\boldsymbol{x} = \boldsymbol{x}^g$, $\boldsymbol{h} = \boldsymbol{h}^{(1)}$, $\boldsymbol{b} = \boldsymbol{b}^g$, $\boldsymbol{c} = \boldsymbol{c}^{(1)}$, $\boldsymbol{W} = \boldsymbol{W}^{(1)g}$, $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{(x)g}$ としている．

分母は

$$\begin{aligned}
&\prod_i \int_{\boldsymbol{x}} \exp\left(-\left(\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_j \frac{W_{ij} x_i h_j}{\sigma_i^2}\right) + \sum_j c_j h_j\right) d\boldsymbol{x} \\
&\qquad = \prod_i \int_{\boldsymbol{x}} \exp\left(-\left(\frac{(x_i^2 - 2x_i b_i + b_i^2)}{2\sigma_i^2} - \sum_j \frac{W_{ij} x_i h_j}{\sigma_i^2}\right) + \sum_j c_j h_j\right) d\boldsymbol{x} \\
&\qquad = \prod_i \exp\left(\left(\frac{-b_i^2}{2\sigma_i^2}\right) + \sum_j c_j h_j\right) \int_{\boldsymbol{x}} \exp\left(\frac{(-x_i^2 + 2x_i b_i)}{2\sigma_i^2} + \sum_j \frac{W_{ij} x_i h_j}{\sigma_i^2}\right) d\boldsymbol{x} \\
&\qquad = \prod_i \exp\left(\left(\frac{-b_i^2}{2\sigma_i^2}\right) + \sum_j c_j h_j\right) \int_{\boldsymbol{x}} \exp\left(-\frac{x_i^2}{2\sigma_i^2}\right) \exp\left(x_i \left(\frac{b_i}{\sigma_i^2} + \sum_j \frac{W_{ij} h_j}{\sigma_i^2}\right)\right) d\boldsymbol{x}
\end{aligned}$$

となる．さらに上式を x について積分して

$$\begin{aligned}
&= \prod_i \exp\left(\left(\frac{-b_i^2}{2\sigma_i^2}\right) + \sum_j c_j h_j\right) \exp\left(\frac{\sigma_i^2}{2}\left(\frac{b_i}{\sigma_i^2} + \sum_j \frac{W_{ij} h_j}{\sigma_i^2}\right)^2\right) \left(\sqrt{2\sigma_i^2 \pi}\right) \\
&= \left(\sigma_i \sqrt{2\pi}\right) \prod_i \exp\left(\frac{1}{2}\left(\sum_j W_{ij} h_j\right)^2 + \sum_j c_j h_j + \frac{b_i W_{ij} h_j}{\sigma_i^2}\right)
\end{aligned}$$

となる．

したがって，

$$\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{h}) &= \frac{\prod_i \exp\left(-\left(\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_j \frac{W_{ij} x_i h_j}{\sigma_i^2}\right) + \sum_j c_j h_j\right)}{\left(\sigma_i \sqrt{2\pi}\right) \prod_i \exp\left(\frac{1}{2}\left(\sum_j W_{ij} h_j\right)^2 + \sum_j c_j h_j + \frac{b_i W_{ij} h_j}{\sigma_i^2}\right)} \\
&= \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\left(\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_j \frac{W_{ij} x_i h_j}{\sigma_i^2}\right) + \sum_j c_j h_j - \left(\frac{1}{2}\left(\sum_j W_{ij} h_j\right)^2 + \sum_j \left(c_j h_j + \frac{b_i W_{ij} h_j}{\sigma_i^2}\right)\right)\right) \\
&= \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\left(x_i - \left(b_i + \sum_j W_{ij} h_j\right)\right)^2\right) \\
&= \prod_i \mathcal{N}\left(x | b_i + \sum_j W_{ij} h_j, \sigma_i^2\right)
\end{aligned}$$

を得る．

また，$\boldsymbol{y}^g$ の条件付き確率分布についても $\boldsymbol{x} = \boldsymbol{y}^g$, $\boldsymbol{h} = \boldsymbol{h}^{(L)}$, $\boldsymbol{b} = \boldsymbol{d}^g$, $\boldsymbol{c} = \boldsymbol{c}^{(L)}$, $\boldsymbol{W} = \boldsymbol{W}^{(L+1)gT}$, $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{(y)g}$ とおくことで同様に導出される．

カテゴリカル分布に従う可視変数の条件付き確率分布

次に，$\boldsymbol{x}^c$ の条件付き確率分布を導出する．条件付き確率分布 $p(\boldsymbol{x}^c|\boldsymbol{h}^{(1)})$ は定義より

$$p(\boldsymbol{x}^c|\boldsymbol{h}^{(1)}) = \frac{p(\boldsymbol{x}^c, \boldsymbol{h}^{(1)})}{p(\boldsymbol{h}^{(1)})} = \frac{p(\boldsymbol{x}^c, \boldsymbol{h}^{(1)})}{\sum_{\boldsymbol{x}^c} p(\boldsymbol{x}^c, \boldsymbol{h}^{(1)})}$$

である．簡単のため $\boldsymbol{x} = \boldsymbol{x}^{(c)}$, $\boldsymbol{h} = \boldsymbol{h}^{(1)}$, $\boldsymbol{b} = \boldsymbol{b}^c$, $\boldsymbol{c} = \boldsymbol{c}^{(1)}$, $\boldsymbol{W} = \boldsymbol{W}^{(1)c}$ として，分母と分子にそれぞれ GCDRM の結合確率分布から $p(\boldsymbol{x}, \boldsymbol{h}) = \exp\{-E(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta})\}/Z$ を代入すると

$$\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{h}) &= \frac{\exp\{-E(\boldsymbol{x}, \boldsymbol{h})\}}{\sum_{\boldsymbol{x}} \exp\{-E(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta})\}} \\
&= \frac{\exp\Big(\sum_i \Big(b_i x_i + \sum_j W_{ij} x_i h_j\Big) + \sum_j c_j h_j\Big)}{\sum_{\boldsymbol{x}} \exp\Big(\sum_i \Big(b_i x_i + \sum_j W_{ij} x_i h_j\Big) + \sum_j c_j h_j\Big)} \\
&= \frac{\prod_i \exp\Big(\sum_j c_j h_j\Big) \exp\Big(b_i s_{i'} + \sum_j W_{ij} s_{i'} h_j\Big)}{\prod_i \sum_{i'} \exp\Big(\sum_j c_j h_j\Big) \exp\Big(b_{i'} s_{i'} + \sum_j W_{i'j} s_{i'} h_j\Big)} \\
&= \prod_i \frac{\exp\Big(b_i s_{i'} + \sum_j W_{ij} s_{i'} h_j\Big)}{\sum_{i'} \exp\Big(b_{i'} s_{i'} + \sum_j W_{i'j} s_{i'} h_j\Big)} \\
&= \prod_i \frac{\exp(s_{i'}) \exp\Big(b_i + \sum_j W_{ij} h_j\Big)}{\sum_{i'} \exp(s_{i'}) \exp\Big(b_{i'} + \sum_j W_{i'j} h_j\Big)}
\end{aligned}$$

となる．ここで，$s_{i'}$ はユニット x_i によって表現される状態である．$p(\boldsymbol{x}|\boldsymbol{h}) = \prod_i p(x_i|\boldsymbol{h})$ であるから，

$$p(x_i = 1|\boldsymbol{h}) = \frac{\exp\Big(b_i + \sum_j W_{ij} h_j\Big)}{\sum_{i'} \exp\Big(b_{i'} + \sum_j W_{i'j} h_j\Big)}$$

また，$\boldsymbol{y}^c$ の条件付き確率分布についても $\boldsymbol{x} = \boldsymbol{y}^c$, $\boldsymbol{h} = \boldsymbol{h}^{(L)}$, $\boldsymbol{b} = \boldsymbol{d}^c$, $\boldsymbol{c} = \boldsymbol{c}^{(L)}$, $\boldsymbol{W} = \boldsymbol{W}^{(L+1)cT}$, $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{(y)c}$ とおくことで同様に導出される．

隠れ変数の条件付き確率分布

最後に，$\boldsymbol{h}^{(1)}$ の条件付き確率分布を導出する．条件付き確率分布 $p(\boldsymbol{h}^{(1)}|\boldsymbol{x}, \boldsymbol{h}^{(2)})$ は定義より

$$p(\boldsymbol{h}^{(1)}|\boldsymbol{x}, \boldsymbol{h}^{(2)}) = \frac{p(\boldsymbol{x}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)})}{p(\boldsymbol{x}, \boldsymbol{h}^{(2)})} = \frac{p(\boldsymbol{x}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)})}{\sum\limits_{\boldsymbol{h}^{(1)}} p(\boldsymbol{x}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)})}$$

である．簡単のため $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{(x)}$ として，分母と分子にそれぞれ GCDRM の結合確率分布から $p(\boldsymbol{x}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) = \exp\{-E(\boldsymbol{x}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}; \boldsymbol{\theta})\}/Z$ を代入すると

$$\begin{aligned}
p(\boldsymbol{h}^{(1)}|\boldsymbol{x}, \boldsymbol{h}^{(2)}) &= \frac{\exp\{-E(\boldsymbol{x}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)})\}}{\sum\limits_{\boldsymbol{h}^{(1)}} \exp\{-E(\boldsymbol{x}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}; \boldsymbol{\theta})\}} \\
&= \frac{\exp\left(-\sum\limits_i \left(\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum\limits_j \frac{W_{ij}^{(1)} x_i h_j^{(1)}}{\sigma_i^2}\right) + \sum\limits_k \left(c_k^{(2)} h_k^{(2)} + \sum\limits_j W_{jk}^{(2)} h_j^{(1)} h_k^{(2)}\right) + \sum\limits_j c_j^{(1)} h_j^{(1)}\right)}{\sum\limits_{\boldsymbol{h}^{(1)}} \exp\left(-\sum\limits_i \left(\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum\limits_j \frac{W_{ij}^{(1)} x_i h_j^{(1)}}{\sigma_i^2}\right) + \sum\limits_k \left(c_k^{(2)} h_k^{(2)} + \sum\limits_j W_{jk}^{(2)} h_j^{(1)} h_k^{(2)}\right) + \sum\limits_j c_j^{(1)} h_j^{(1)}\right)} \\
&= \frac{\prod\limits_j \exp\left(-\left(\sum\limits_i \frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum\limits_k c_k^{(2)} h_k^{(2)}\right)\right) \exp\left(\sum\limits_i \frac{W_{ij}^{(1)} x_i h_j^{(1)}}{\sigma_i^2} + \sum\limits_k W_{jk}^{(2)} h_j^{(1)} h_k^{(2)} + c_j^{(1)} h_j^{(1)}\right)}{\prod\limits_j \sum\limits_{h_j^{(1)} \in \{0,\ 1\}} \exp\left(-\left(\sum\limits_i \frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum\limits_k c_k^{(2)} h_k^{(2)}\right)\right) \exp\left(\sum\limits_i \frac{W_{ij}^{(1)} x_i h_j^{(1)}}{\sigma_i^2} + \sum\limits_k W_{jk}^{(2)} h_j^{(1)} h_k^{(2)} + c_j^{(1)} h_j^{(1)}\right)} \\
&= \prod_j \frac{\exp\left(\sum\limits_i \frac{W_{ij}^{(1)} x_i h_j^{(1)}}{\sigma_i^2} + \sum\limits_k W_{jk}^{(2)} h_j^{(1)} h_k^{(2)} + c_j^{(1)} h_j^{(1)}\right)}{\sum\limits_{h_j^{(1)} \in \{0,\ 1\}} \exp\left(\sum\limits_i \frac{W_{ij}^{(1)} x_i h_j^{(1)}}{\sigma_i^2} + \sum\limits_k W_{jk}^{(2)} h_j^{(1)} h_k^{(2)} + c_j^{(1)} h_j^{(1)}\right)} \\
&= \prod_j \frac{\exp\left(h_j^{(1)} \left(\sum\limits_i \frac{W_{ij}^{(1)} x_i}{\sigma_i^2} + \sum\limits_k W_{jk}^{(2)} h_k^{(2)} + c_j^{(1)}\right)\right)}{\sum\limits_{h_j^{(1)} \in \{0,\ 1\}} \exp\left(h_j^{(1)}\right) \exp\left(\sum\limits_i \frac{W_{ij}^{(1)} x_i}{\sigma_i^2} + \sum\limits_k W_{jk}^{(2)} h_k^{(2)} + c_j^{(1)}\right)}
\end{aligned}$$

となる．ここで，$p(\boldsymbol{h}^{(1)}|\boldsymbol{v}, \boldsymbol{h}^{(2)}) = \prod_j p(h_j^{(1)}|\boldsymbol{v}, \boldsymbol{h}^{(2)}),\ h_j^{(1)} \in \{0,\ 1\}$ であるから，

$$\begin{aligned}
p(h_j = 1|\boldsymbol{v}, \boldsymbol{h}^{(2)}) &= \frac{\exp\left(\sum\limits_i \frac{W_{ij}^{(1)} x_i}{\sigma_i^2} + \sum\limits_k W_{jk}^{(2)} h_k^{(2)} + c_j^{(1)}\right)}{1 + \exp\left(\sum\limits_i \frac{W_{ij}^{(1)} x_i}{\sigma_i^2} + \sum\limits_k W_{jk}^{(2)} h_k^{(2)} + c_j^{(1)}\right)} \\
&= \sigma\left(c_j^{(1)} + \sum_i \frac{W_{ij}^{(1)} x_i}{\sigma_i} + \sum_k W_{jk}^{(2)} h_j^{(1)} h_k^{(2)}\right)
\end{aligned}$$

を得る.

また, $\boldsymbol{h}^{(L)}$ の条件付き確率分布についても $\boldsymbol{x} = \boldsymbol{y}$, $\boldsymbol{h}^{(1)} = \boldsymbol{h}^{(L)}$, $\boldsymbol{h}^{(2)} = \boldsymbol{h}^{(L-1)}$, $\boldsymbol{b} = \boldsymbol{d}$, $\boldsymbol{c}^{(1)} = \boldsymbol{c}^{(L)}$, $\boldsymbol{c}^{(2)} = \boldsymbol{c}^{(L-1)}$, $\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(L+1)T}$, $\boldsymbol{W}^{(2)} = \boldsymbol{W}^{(L)T}$, $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{(y)}$ とおくことで同様に導出される.

B　音声認識・合成実験に使用した音素記号の一覧

vowels	/a/, /i/, /u/, /e/, /o/, /A/, /I/, /O/, /U/
consonants	/k/, /s/, /t/, /ts/, /n/, /h/, /f/, /m/, /r/, /g/, /z/, /j/, /d/, /b/, /p/, /ky/, /sh/, /ch/, /ny/, /hy/, /my/, /ry/, /gy/, /dy/, /by/, /py/
semivowels	/y/, /w/
special moras	/N/, /cl/, /pau/
others	/xx/, /sil/

特殊モーラにおける音素記号「/N/」は撥音「ん」を,「/cl/」は促音「っ」を,「/pau/」は文節などに生じる無声区間を表す．その他の音素記号「/xx/」は未知の音素を表す．また,「/sil/」は音声ファイルの中で，実際に発話が開始される前と後に生じる無声区間を表す.

C　音声認識・合成実験に使用したコンテキストラベルの一覧

the difference between accent type and position of the current mora identity	−49 ~ 49
position of the current mora identity in the current accent phrase (forward)	1 ~ 49
position of the current mora identity in the current accent phrase (backward)	1 ~ 49
pos (part-of-speech) of the previous word	
inflected forms of the previous word	
conjugation type of the previous word	
pos (part-of-speech) of the current word	
inflected forms of the current word	
conjugation type of the current word	
pos (part-of-speech) of the next word	
inflected forms of the next word	
conjugation type of the next word	
the number of moras in the previous accent phrase	1 ~ 49
accent type in the previous accent phrase	1 ~ 49
whether the previous accent phrase interrogated or not (0: not interrogative, 1: interrogative)	
undefined context	
whether pause insertion or not in between the previous accent phrase and the current accent phrase	
the number of moras in the current accent phrase	1 ~ 49
accent type in the current accent phrase	1 ~ 49
whether the current accent phrase interrogated or not (0: not interrogative, 1: interrogative)	
undefined context	
position of the current accent phrase identity in the current breath group by the accent phrase (forward)	1 ~ 49
position of the current accent phrase identity in the current breath group by the accent phrase (backward)	1 ~ 49
position of the current accent phrase identity in the current breath group by the mora (forward)	1 ~ 49
position of the current accent phrase identity in the current breath group by the mora (backward)	1 ~ 49
the number of moras in the next accent phrase	1 ~ 49
accent type in the next accent phrase	1 ~ 49
whether the next accent phrase interrogated or not (0: not interrogative, 1: interrogative)	
undefined context	
whether pause insertion or not in between the next accent phrase and the current accent phrase	
the number of accent phrases in the previous breath group	1 ~ 49
the number of moras in the previous breath group	1 ~ 99
the number of accent phrases in the current breath group	1 ~ 49
the number of moras in the current breath group	1 ~ 99
position of the current breath group identity by breath group (forward)	1 ~ 19
position of the current breath group identity by breath group (backward)	1 ~ 19
position of the current breath group identity by accent group (forward)	1 ~ 49
position of the current breath group identity by accent group (forward)	1 ~ 49
position of the current breath group identity by mora (forward)	1 ~ 199
position of the current breath group identity by mora (forward)	1 ~ 199
the number of accent phrases in the next breath group	1 ~ 49
the number of moras in the next breath group	1 ~ 99
the number of breath groups in this utterance	1 ~ 19
the number of accent phrases in this utterance	1 ~ 49
the number of moras in this utterance	1 ~ 199
the number of frames in the current phoneme	
the relative position of the current frame in the current phoneme	